Computational studies on recurrent genomic selection for genetic improvement of soybeans

by

Vishnu Ramasubramanian

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:

William Beavis, Co-major Professor
Alicia Carriquiry, Co-major Professor
Jack Dekkers
Karin Dorman
Lizhi Wang

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/ thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2021

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I would like to dedicate this dissertation to my parents and colleagues at Iowa State University. I can't thank Dr. Beavis and his lab members enough for their constant support. I need to thank Dr. Beavis for his constant support in the midst of all kinds of challenges. Through all these years, I've been with his research group, he has always been an extremely welcoming person and patient with everything. I could not have done anything without Dr. Beavis' guidance at every stage of this dissertation project, as Applied Plant Breeding and Quantitative Genetics were mostly a new field of study for me. He was also extremely patient in explaining all the concepts and encouraged independent work. I would also like to thank my POS committee members Dr. Alicia Carriquiry, Dr. Jack Dekkers, Dr. Karin Dorman and Dr. Lizhi Wang for their helpful suggestions and guidance throughout the dissertation project.

I must also thank the research group members Danielle, John, Haley, Reka, and Bongsong. I've never seen colleagues more patient and understanding than this group and they often worked hard to make it a welcoming team. I must also thank the larger ISU community for their support and making ISU a great campus town. Their engagement and involvement in affairs of the campus and larger community completely changed my perception of life in a community. Having lived with only small groups of friends and family with very little direct engagement with the larger community, I found campus life in Ames to be a completely different experience that provided a sense of community engagement. I believe this experience will stay with me as long as I can remember things. Thanks to the ISU community and the larger Ames community for that.

I would also like to thank my parents for giving me the freedom to choose my career and a scholarly way of life. Their support has been a constant source of encouragement since high

school to pursue higher education. I would also like to thank my cousins and their families for encouraging me to pursue higher education in the US.

ABSTRACT

Every crop genetic improvement project has unique objectives, although there are some general consistent objectives that are common among projects. Due to competition in the marketplace commercial plant breeders need to place greater emphasis on immediate genetic gains with potential loss of useful genetic variance. In contrast public plant breeders have the opportunity, perhaps obligation, to emphasize retention of useful genetic variance while improving genetic improvements. Most often the relative emphasis on these competing objectives has not been designed rather it has emerged as a consequence of reproductive biology, genetic architecture and budget constraints. The theme of the research reported in this dissertation is that the development of Genomic Selection (GS) methods has provided the ability to plan and execute the trade-offs between these competing objectives in genetic improvement projects.

Some early simulation studies that compared recurrent GS with Phenotypic selection (PS) revealed greater short term genetic gains with GS, while PS resulted in better long term genetic gains because PS retains useful genetic variability during the early cycles of selection. These early studies also indicated different genetic architectures, heritabilities, GS methods, training sets and selection intensities would result in a range of responses across multiple cycles. We hypothesized that interactions among these factors could further increase the number of possible response curves. We decided to evaluate the hypothesis by simulating 40 cycles of recurrent selection for sets of founders with genotypic data and population structures based on the founders of the Soybean Nested Association Mapping panel. Ten simulations were conducted on a factorial set comprised of over 300 combinations of selection methods, training sets, and

selection intensities, which are under the control of the breeder, as well as genetic architecture and heritability, which are not.

To distinguish among the 300+ replicated response curves we employed a first order recurrence equation to model the genotypic responses. Because recurrence equations are discrete analogs of differential equations, the estimated parameters enabled evaluation of response rates, half-lives and genotypic values as responses approach asymptotic limits. By modeling genotypic responses it was also possible to conduct ANOVA of the non-linear responses, which revealed that both the rates of genetic improvement in the early cycles and limits to genetic improvement in the later cycles are significantly affected by interactions among all investigated factors. Even though all possible interactions significantly affected modeled responses, there were some consistent trends. Updating GP models with training sets consisting of data from prior cycles of selection significantly improved prediction accuracy and genetic response for all GS methods. From among the GS methods with updated training sets, selection on values estimated from Ridge Regression –Restricted Maximum Likelihood Method (RR-REML) resulted in better response rates and larger asymptotic limits than selection on estimates from BayesB and Bayes LASSO models. A Support Vector Machine with a radial basis kernel method resulted in the fastest loss of genetic variance in the early cycles.

We next hypothesized that we could improve both response rates and retention of useful genetic variability in the simulated soybean populations by decomposing breeding strategies into decisions about selection methods and mating designs. For breeding populations organized into islands, decisions about possible migration rules among family islands were included. From among 60 possible strategies, genetic improvement is maximized for the first five to ten cycles using GS and a hub network mating design in breeding populations organized as fully connected

family islands and migration rules allowing exchange of two lines among islands every other cycle of selection. If the objectives are to maximize both short-term and long-term gains, then the best compromise strategy is similar except a genomic mating design, instead of a hub networked mating design, is used. This strategy also resulted in realizing the greatest proportion of genetic potential of the founder populations. In Weighted Genomic Selection (WGS), the estimated marker effects are weighted by the inverse of the favorable allele frequency and the weighted values are used for selection. WGS, when applied to both non-isolated and island populations, also resulted in the realization of the greatest proportion of genetic potential of the founders, but required many more cycles than the strategy that showed the best compromise between short-term and long-term gain in the case of non-isolated populations. These studies have the potential to contribute to the development of decision support systems that use new approaches to integrate the strengths of whole-genome level information, prediction modeling, and optimization methods for long-term genetic improvement of crops.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Genetic improvement of cultivated crops is the core purpose for plant breeding. People have been remarkably successful in changing the desirable traits of crops to suit the demands of a growing population by applying informed methods of artificial selection over several thousand years. Modern Plant breeding programs consist of 1) recurrent genetic improvement projects, 2) variety development projects 3) trait introgression projects and 4) product placement projects (Fehr, 1991). Past genetic improvements are assessed using realized genetic gain, which is an estimate of change of the average genotypic value for traits across recurrent cycles of selection and inter-mating. Perhaps the most relevant trait for assessment of realized genetic gain is yield per unit land. Of concern are gains in many crop species that have stagnated (Bhatia et al. 2008; Van Ittersum et al. 2013; Liu et al. 2016).

Estimates of genetic gain combined with estimated yield potential provide an assessment of genetic improvement relative to the maximum possible. Studies on soybean variety yields over the past 80 years in North America have shown a trend of increase. Estimates of yield components suggest that two-thirds of yield gain is due to genetic improvement and one-third due to improved agronomic practice (Mikel et al. 2010; Rincker et al. 2014). Given the improvements that have been made, more sophisticated methods for selection will be needed to achieve further improvements in yield (Hickey et al. 2017).

Several directions of research were pursued over the past century to improve selection methods (Hickey et al. 2017). Concepts such as breeding value allowed breeders to estimate the value of any individual line as a parent based on the predicted phenotypic values of progeny. Advancements in genotyping and molecular biology methods led to the use of marker assisted

selection (MAS), where genetic markers that are associated with improved trait values are considered identifiers for lines with high trait values (Hickey et al. 2017). This requires linkage between marker alleles and causative alleles as well as the ability to identify markers with significant associations between polymorphic markers and trait variability. However the estimation of significance of markers has complications due to 'Beavis Effect', which states that estimates of genetic effects associated with the significant marker alleles are inflated (Beavis 1994). Several methods were proposed to counter this over-estimation by shrinking the estimates. In 2001, Meuwissen et al. proposed statistical methods to estimate genotypic values of individuals based on genome-wide marker sets. These statistical algorithms are referred to as genomic prediction (GP) methods. GP methods obviate any need for identification of specific marker-trait associations, a.k.a., quantitative trait loci (QTL). Meuwissen's approach motivated development of a large number of methods to predict genomic estimated breeding values (GEBV) of genotypes (Bernardo 2008, 2014; Jannink et al. 2010; Asoro et al. 2011; Heslot et al. 2012; Nakaya and Isobe 2012; Hagan et al. 2012 ; Emily and Bernardo 2013; Crossa et al. 2014; Heslot et al. 2015; Liu et al. 2015; Beyene et al. 2015; Bassi et al. 2016; Marulanda et al. 2016; Jonas and de Koning 2013, 2016; Hickey et al. 2017; Goiffon et al. 2017). GP methods have also been compared in terms of prediction accuracy and mean squared error, which are the two most commonly used evaluation metrics (Long et al. 2010, 2011; Guo et al. 2012; Howard et al. 2014). In plant breeding, GS methods have been adopted by commercial breeding organizations with well-funded R&D programs and a few public breeding programs with sufficient financial resources to pay for data from high throughput genotyping and field trials (Guzman et al. 2016; Hickey et al. 2017). For example, the CIMMYT wheat improvement program uses a combination of PS and GS, where GS is used only for filtering lines that have been discarded in

the initial selection steps (Guzman et al. 2016). Quantitative geneticists have investigated GP using simulations to compare genetic gain in recurrent selection programs. By directly selecting for markers that are associated with traits, GS methods achieve rapid increases in genetic gain compared to PS phenotypic selection. However, PS does not lose genetic diversity as fast and results in less limited responses to selection in closed populations.

Motivation

Soybean represents an extreme example of a self-pollinated crop. Due in part to labor intensive and expensive procedures required to successfully cross pollinate soybeans [NCSRP], there is limited genetic diversity among soybean varieties adapted to production environments (Mikel et al. 2010). Production environments are subdivided into maturity zones. In North America, soybean varieties adapted to maturity zone IV will not mature in time to harvest if planted in maturity zone II. These MZ's further subdivide breeding populations and limit genetic diversity of the breeding populations from which varieties are derived. How rapidly will genetic gains approach a limit using limited genetic diversity among soybean accessions adapted to maturity zones? How fast will the genetic potential of a breeding population be lost due to phenotypic and genomic selection methods, genetic architectures, and selection intensities?

Factors such as genetic architecture, marker density, training population size and structure affect prediction accuracy and responses in recurrent selection. While these factors have been previously identified with simulations, combinations of these factors and their interactions have not been systematically investigated in the context of recurrent GS to examine their impact on short-term and long-term responses.

In an ideal scenario it should be possible to realize the maximum possible genetic potential from any genetic value in the current generation in just one cycle of selection (Figure

1). However it is not feasible to allocate astronomically huge amounts of resources to evaluate all possible combinations of crosses to identify the crosses that will result in maximal genetic gain. A realistic feasible option is to maximize genetic gain in one cycle among a subset of all possible crosses. However, such maximization of genetic gain will result in loss of genetic variance and genetic potential in the population (Figure 1). Such loss of genetic potential is mostly due to the loss of favorable alleles in the population due to drift and loss of favorable alleles during the selection process due to linkage or linkage disequilibrium with unfavorable alleles. The other extreme is to maximize retention of genetic potential with minimal genetic gain, which is contrary to the main objective of genetic improvement project (Figure 1). In between these two extremes, there are an infinite number of unique curves that represent unique tradeoffs between maximizing genetic gain and maximizing retention of genetic potential. In order to identify response curves that represent optimal tradeoffs between maximizing genetic gain in the short-term and retaining genetic potential in the population for long-term gains, we've split the recurrent selection process into distinct non-overlapping steps of prediction, selection and crossing and investigated factors that impact each of these steps.

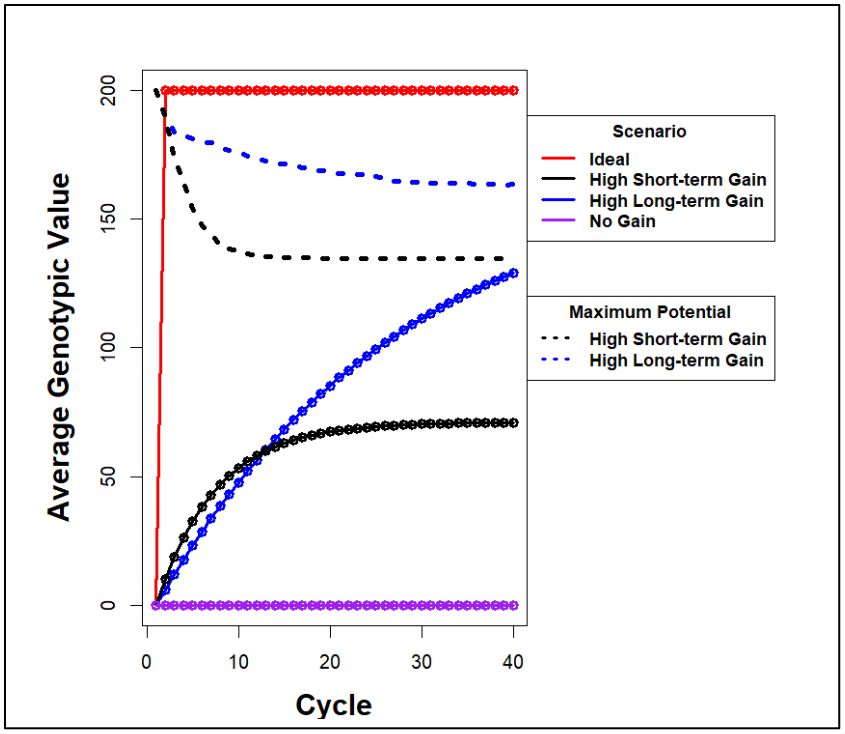

Figure 1 Illustration of response to recurrent selection in four scenarios including i) ideal (red), ii) response pattern with high short-term gain and low limit (black), iii) response pattern with high long-term gain (blue), iv) no gain (purple). The dashed lines represent maximum genotypic potential available in the population for i) high short-term gain (black), ii) high long-term gain (blue).

In the following dissertation chapters, we've attempted to answer the following specific questions. Do GP models used to train impact responses in the short-term and long-term? Do GP models differ in rate of decay of prediction accuracy? How to prevent reduction in accuracies using retraining of GP models to realize greater responses? How training sets used to update GP models impact short-term and long-term responses and how it interacts with other factors? How accuracy and responses in recurrent GS are impacted by factors such as genetic architecture, heritability of traits, selection intensity?

In addition to prediction accuracy to GP models, there are other factors such as loss of genetic variance and loss of favorable alleles in closed systems that impact limits of responses. How to deal with these constraints? While the recommended approaches for dealing with constraints include adjusting selection intensity and weighted genomic selection (WGS), in practice, breeders exchange lines among breeding stations that have similar conditions for cultivation. This process is driven by observation and performance of individual genotypes and expectations of short-term gain and not on expectations for performance of future generations and long-term gain.

A plant breeder has to make many decisions regarding this exchange process as to what lines to consider and which breeding station to consider for exchange. How many breeding stations to consider and what should be the nature of connections among these stations? How many genotypes should be exchanged and from where should these genotypes be acquired? We've modeled this exchange process in recurrent selection using an island model evolutionary algorithm that takes into account all these factors. We've also compared responses from global selection methods with local selection methods with and without exchange of genotypes.

In addition to truncation selection employed in many proposed GS methods, there are other multi-objective optimization based selection methods, which offer optimal solutions considering the many objectives of plant breeding programs. For example, an objective could be to maximize genetic gain while maintaining a defined threshold for genetic variance in the population. We have compared one such method in the context of global and local selection.

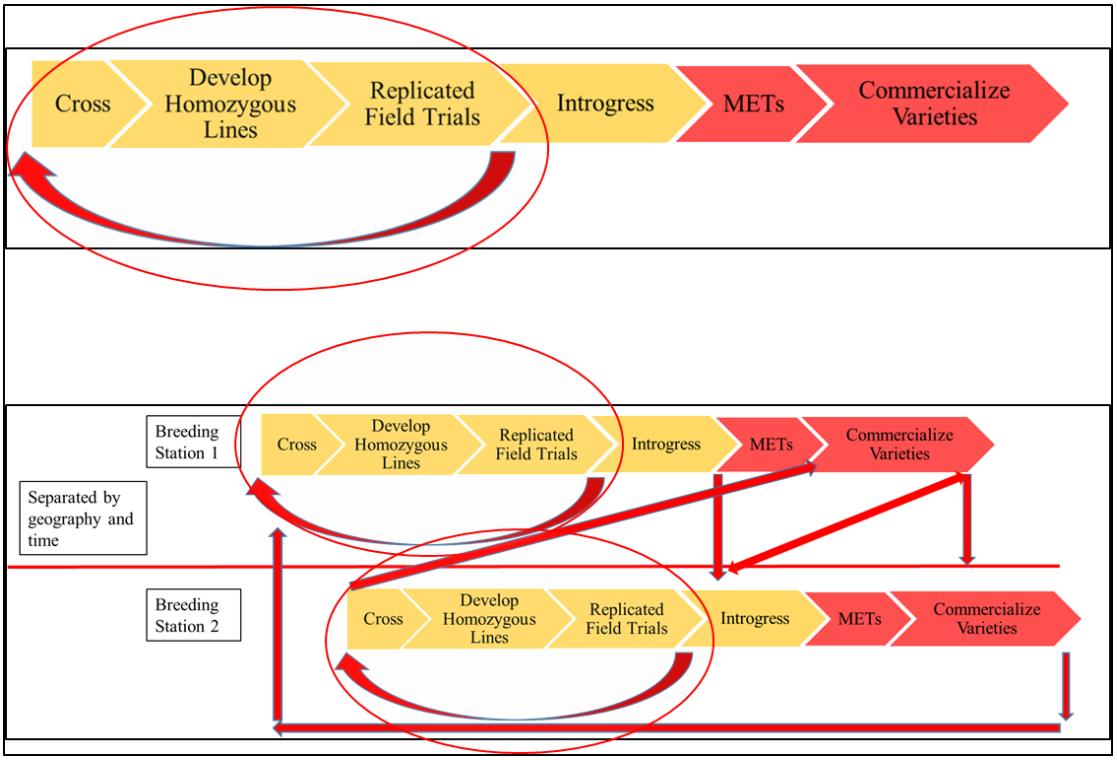

Figure 2 Schematic of cultivar development pipeline. Upper panel shows pipeline employed in one breeding station. This process is simulated in most theoretical studies of recurrent selection. Lower panel depicts a similar cultivar development pipeline employed in multiple breeding stations with exchange among breeding stations. The first three steps inside the elliptical region comprise the recurrent selection or population improvement process with generation, evaluation, selection, crossing steps that are iterated multiple times. The arrows point to transfer or exchange of crossing material within and among plant breeding stations.

Because breeding islands are analogous to soybean breeding programs consisting of multiple breeding sites (islands), we've identified several populations, where Island Genomic Selection (IGS) and Island Weighted Genomic Selection (IWGS) have the potential to improve genetic gains using simulated conditions that approximate public soybean breeding in Maturity

Zone III of the United States. For these simulations, data from USDA-Germplasm Collection and Soybean Nested Association Mapping panel was used as a model. In the following sections of this introductory chapter, we have briefly described the concepts and methods used.

1.2 Genetic Gain

Yield is the most important trait for producers of commodity soybeans. Since these farmers represent over 95% of the market plant breeders also emphasize genetic gains for yield. Yield is measured in bushels/acre (bu/ac) or kilogram/hectare (Kg/ha). As the units suggest, it is considered an estimate of productivity per unit of land, with crop genotypes, environment and agronomic practices used to realize the productivity. Yield potential refers to the maximum yield attainable for a genetic variety given perfect environmental and management inputs. Yield potential of a variety could vary depending on variation in environmental conditions such as solar radiation, water availability, fertilizers and pest management. Yield gap refers to the difference between yield potential of a variety and attained yield in any given growing season (Cassman et al. 2003; Bhatia et al. 2008; Lobell et al. 2009; Van Ittersum et al. 2013; Liu et al. 2016).

Realized genetic gain for yield refers to changes over time. Since it assumes perfect environmental and management factors, genetic improvement is specifically an estimate of the genetic component (Mikel et al. 2010; Rincker et al. 2014; Hickey et al. 2017; Byrum et al. 2017). This is contrasted with yield protection which allows farmers to bridge the yield gap between yield potential and realized yields of cultivated genotypes.

Obtaining an accurate estimate of genetic gain is complicated, as it involves separating the genetic and non-genetic components. Some estimates of genetic gain such as the one in the breeder's equation (2), assumes constant environmental and management inputs. Byrum et al

(2017) derived an estimate called Genetic Gain Performance measure that uses advanced analytics to separate the genetic from non-genetic components. Applying this estimate in operational decisions led to a significant improvement in genetic gain in Syngenta Soybean Breeding Program (Byrum et al. 2017).

The genetic potential is the maximum possible genotypic value (assuming no environmental variation) that can be achieved with the combination of all favorable alleles that contribute to the trait. A limit to genetic gain indicates that there can be no further genetic improvement without introducing useful genetic variability into the breeding population. Limits of responses to selection are usually less than the genetic potential of the founders of a breeding population. Most breeding populations of crop genotypes respond to selection but at levels well below the initial genetic potential of the founders.

Breeder's Equation

The breeder's equation has been used to predict genetic response using estimates of heritability and selection differential. There are univariate and multivariate forms of the breeder's equation. The different forms of the equation partition genetic response into two components that are due to selection intensity, the estimated phenotypic standard deviation and the estimate of heritability. The breeder's equation could be considered a special condition of Price's equation (Queller 2017) when the correlation between breeding value and fitness is zero, which means variance in phenotypic trait values completely capture the variance in fitness and there is no other correlated genetic effect (Okasha 2008; Morrissey et al. 2010; Pitchers et al. 2014; Queller 2017).

$$\mathbf{R} = \frac{\mathbf{ir\sigma_A}}{\mathbf{t}} \quad \textbf{(1)}$$

R - response to selection
i - selection intensity
r - selection accuracy
σ_A - s.q.r.t. of additive genetic variance
t - time in years between crossing
and selection of new progeny genotypes

Equation (1) expresses genetic response in terms of the selection intensity, additive genetic variance, accuracy of selection and time required to complete a cycle of selection and creating a new set of selectable progeny. Selection accuracy is equivalent to prediction accuracy when selecting on genotypic values. 't' in the denominator denotes the time, in terms of years, that is required between creating a set of genotypes and selecting those genotypes which will create a new set of genotypes. Often the form used by breeders includes a cost factor in the denominator that allows them to estimate genetic gain per unit of investment and cycle interval (Equation 2).

$$\mathbf{R} = \frac{\mathbf{ir\sigma_A}}{(\mathbf{cost} * \mathbf{t})} \quad \textbf{(2)}$$

R - response to selection
i - selection intensity
r - selection accuracy
σ_A - s.q.r.t. of additive genetic variance
t - time in years between crossing
and selection of new progeny genotypes
cost - cost of selection program per generation

Equation (3) expresses genetic response in terms of selection differential and estimated narrow sense heritability.

$$\mathbf{R = Sh^2} \quad \textbf{(3)}$$

S - selection differential

h^2 - narrow sense heritability

Multivariate versions of Breeder's equation are given in equation (4). An important assumption in breeder's equation is normality of trait values. Another assumption is causation, where all the other traits that have a causal role in contributing to trait values are included in the model. It means that all correlated traits that respond to selection must be included in the model as co-variates for the prediction to be accurate. In situations where there is phenotypic correlation among traits and no genetic correlations, the correlated traits won't respond to selection. When there is negative correlation, an increase in one trait due to directional selection will decrease the trait that is negatively correlated (Okasha 2008; Morrissey et al. 2010; Pitchers et al. 2014; Queller 2017).

$$\mathbf{R = GP^{-1}S} \quad \textbf{(4)}$$

R - vector of response elements

G - variance - covariance matrix of genetic values

P - variance - covariance matrix of phenotypic values

S - vector of selection differentials

Application of Breeder's Equation

The breeder's equation is an example of *ceteris paribus* law. A *ceteris paribus* law ('given all things equal') states that the relation among variables described by the law holds true only for the defined set of conditions or constraints (Reutlinger et al. 2019). So there are limitations to the application of Breeder's equation despite its widespread use in predicting genetic gains. For example, the breeder's equation may not predict outcomes from long-term selection because heritabilities in subsequent cycles of selection will change as the genetic

variance/covariance will change depending on the population structure among founders, genetic architecture of the traits and selection intensity. Assuming an infinitesimal model for genetic architectures of quantitative traits, in which a large number of alleles with small effects are responsible for trait values, genotypic and phenotypic values will be normally distributed, in the absence of selection (Barton et al. 2017). However, selection introduces bias in the distribution of phenotypic values in future generations and violates normality assumptions required for the relationship in Breeder's equation to hold true (Pigliucci 2006; Blows and Walsh 2009).

The Breeder's equation also has limited applicability to predict response in natural populations as covariance between fitness values and genotypic values are influenced by variable non-genetic factors. Thus, it is widely accepted that Robertson-Price theorem of natural selection has more general application for predicting the evolution of natural populations (Okasha 2008; Morrissey et al. 2010; Queller 2017).

The breeder's equation has also been applied in the context of GS, where selection accuracy corresponds to prediction accuracy between training and validation sets, intensity of selection adjusted for GS and additive genetic variance remain the same as in the Breeder's equation. Given the limitations of the breeder's equation to predict long-term evolutionary dynamics of populations under artificial selection, simulation studies have been used to investigate response to selection across many selection cycles.

1.3 Simulation Studies in Plant Breeding

Initial modeling studies that attempted to find predictive relationship between phenotype and genotype relied on analytical treatments on simple systems with few loci and no interaction among those loci (Kempthorne, 1969). While the simple assumptions yielded good theoretical foundation for linking Mendelian genetics with Darwin's evolutionary theory, the analytical

treatment failed to account for the complexity of evolutionary dynamics of real populations. Given the limits of analytical approaches to address the complexity of experimental systems, simulations have been used to identify factors and breeding strategies that achieve desired responses of t recurrent selection programs (Jannink 2010; Liu et al. 2015; Akdemir and Sánchez 2016; Yabe et al. 2016; Akdemir et al. 2019).

The application of approximate genotype-to-phenotype (G-to-P) mapping functions enables reasonable simulated phenotypic values from genotypic data. It is also possible to simulate genomes and biological processes like selection, meiotic recombination and choice of mates to investigate responses of breeding populations *in silico* in a few hours to a few days. Even though simulations use models that are abstract approximations of the true underlying biological mechanisms, their use has provided reasonable understanding of dynamic patterns of responses to selection in domesticated and natural populations. In addition, there are ambitious attempts to use big data and simulation modelling for precision breeding. Precision breeding is a term that describes selection of mates that will provide optimal genetic improvement (Podlich and Cooper, 1999; Cooper et al. 2002; Cooper et al. 2014).

To date, mapping biological mechanisms of G-to-P are incomplete and poorly understood, even with all the genomic, network and biochemical pathway discoveries that have been made in the last 30 years. Current methods treat the processes and multiple levels of organization from G-to-P as black boxes, although they often demonstrate high prediction accuracies *in vitro* genetic improvement projects. Cooper et al (2002) classified quantitative traits that are of interest to plant breeders based on heritability and genetic complexity and suggested widely different modeling strategies to evaluate response to selection. In their classification scheme, complexity refers to the genetic architecture of traits, characterized by the

number of genetic loci that regulate the trait, the distribution of allelic effects and the presence of gene-by-gene and gene-environment interactions. The accuracy of prediction can be low even for simple traits with a small number of genes contributing the trait when the heritability of the trait is low.

Factorial Design for Evaluations of Responses to Selection

Previous investigations of responses to genomic selection investigated single factors, but to our knowledge did not identify the influences of interactions among factors. To investigate multiple combinations of factors that have been shown to individually influence response to selection, we used factorial designs. The factorial design provided both efficient estimation of main effects and as well as information on responses to interactions of factors. The ANOVA of linear or non-linear models are used to identify significant combinations of factors that affect the response variables (Myers 1976; Collins et al. 2014; Dunn 2020). This makes factorial design is optimal for selecting the combination of factors among the extreme scenarios (Figure 1) that will result in response curves that meet the objectives of the recurrent selection program.

1.4 Recurrent Genomic Selection

Three general approaches have been used to investigate improvement of genetic gain with GS. The first approach is through development of GP models, where improving the accuracy of the GP model is expected to improve genetic gain by improving the accuracy of selection (Habier et al. 2007; Goddard 2009; Zhong et al. 2009; Jannink 2010; Heffner et al. 2011; Bastiaansen et al. 2012; Bijma 2012; Wimmer et al. 2013; Lorenz 2013; Hickey et al. 2014, 2017). The second approach addresses selection of lines as parental lines for creating a new set of progeny for evaluation (Cochran 1951; Bertan et al. 2007; Bos and Caligari 2008;

Bernardo 2014). The third approach considers various crossing strategies, which affects the utilization of selected individuals to achieve specific goals in subsequent cycles of selection and mating (Akdemir & Sánchez 2016; Xu et al. 2017; Goiffon et al. 2017; Gorjanc et al. 2018).

Prediction accuracies of GP models employed in recurrent selection are affected by factors such as genetic architecture of traits, marker density, composition and number of genotypes included in training sets, and GP model. Prediction accuracy decays with selection as the genetic variance-covariance structure of populations change with over recurrent cycles of selection and crossing. This decay of prediction accuracy reduces the limits of responses that are achieved, but it can be prevented or reduced by retraining GP models with updated training sets. In chapter 2 of this dissertation, we have described the impact of all these factors and their interactions on prediction accuracy and response to selection.

Trade-offs in Genetic Improvement

Plant breeders always need to address more than one objective. Often the breeding objectives are related through reciprocal and negative associations (Saeki et al. 2014). However, it is not true that such negative associations will prohibit optimization of multiple objectives, where optimality is defined as the best compromise among competing criteria. Dewitt and Langerhans (2004) identified four strategies that organisms use to minimize trade-offs among competing objectives. These strategies include specialization in which genotypes are well adapted to specific environments. Generalists are genotypes with moderate fitness in many environments. Bet-hedging refers to a strategy where there is switch between genotypes at some probability and phenotypic plasticity where genotypic switches are based on environmental triggers. Constraints such as genetic, functional, developmental and historic contingency prevent the evolution of optimal genotypes. However plant breeders aim to create genotypes that are

optimally adapted to growing conditions encountered in a targeted population of environments (TPE) and consequently minimize the trade-offs between specializations, generalization, bet hedging vs phenotypic plasticity. The “law of diminishing returns” poses another constraint, where improving traits to 80% of the potential will require only 20% as much effort as required to realize the genetic potential. In addition to high costs, there are short-term and long-term tradeoffs among economic, environmental and social factors in adoption of high yielding varieties (Logan 2017).

Genetic Constraints in Crop Improvement

Genetic constraints include loss of genetic variance in the population and loss of beneficial alleles that impact realization of maximum potential in later cycles. Selection for domestication and crop improvement often leads to depletion of genetic variability in the population. This has prompted some to suggest introduction of genotypes from geographically separate regions or from historical relatives, e.g., landraces, through crossing, gene stacking or other application of biotech methods such as genetic engineering or gene editing (Garcia 2013; Gorjanc et al. 2016; Bailey-Serres et al. 2019; Allier et al. 2019). Simulation studies of strategies that involve a combination of GS and genome editing called PAGE (Promotion of Alleles by Genome Editing) demonstrated step improvements compared to implementations of GS alone (Janez et al. 2015; Hickey et al. 2016, 2017).

One of the common approaches to address constraints in recurrent PS & GS programs is to adjust selection intensity, where a higher intensity of selection is used for rapid short term genetic gains (GS) and relaxed selection intensities for greater long-term gain at a slower rate (PS). Another approach is to implement strategies that allow informed selection and crossing decisions for long-term genetic improvement of their crops with minimal trade-

offs for short-term gains. For example, in WGS (weighted genomic selection), rare favorable alleles with high marker effects are given greater weights to retain alleles in the selection population and maintain genetic variance through more cycles of recurrent selection (Jannink 2010; Liu et al. 2015).

In practice, plant breeders tend to exchange genotypes among breeding stations that have similar environmental conditions (Figure 1). While the global collection of genotypes that could be used is large, plant breeders often tend to select high yielding lines that are well adapted to environmental conditions in their TPE. Even for development of new cultivars, plant breeders tend to be selective about lines that are adapted to other geographical regions as it takes a significant portion of limited resources any given year to develop and evaluate new cultivars (personal communication from Andrew Scaboo, MO, Brian Diers, Ill, Bill Schapaugh, KSU, and Asheesh Singh, ISU). It is not clear if the practice of limited exchange of genotypes among soybean breeders is informed by theory or simply a desire to take advantage of the best available genotypes for the TPE.

Since Wright introduced the concept of adaptive landscapes for his model of evolution in shifting balance theory, a number of researchers have extended and applied this concept to artificial selection (Wright 1932; Wright 1988; Cooper 2002; Yabe et al. 2016). These evolutionary concepts were appropriated by computer scientists to solve complex optimization problems. Solutions to these multi-objective optimization problems were obtained by developing algorithms that are referred to as evolutionary algorithms (Goldberg 1989; Goldberg and Deb 1992). Genetic algorithms (GA) are a sub-class of evolutionary computing algorithms of particular interest because they are based on artificial recurrent selection. The evolution of populations can be considered a heuristic optimization process, where the individual members of

the populations are solutions to adaptive challenges. Any individual's adaptation to its specific challenge can be captured in an individual's fitness value estimated with a function that maps genotypes to fitness with phenotypic values as an intermediate variable. In artificial selection, phenotypic values such as yield can be considered equivalent to fitness values as selection is based on phenotypic values of the selection units. Selection units can be individuals or in the case of crop species, replicable genotypes such as lines, varieties, hybrids, clones, etc. The genetic variation of selection units and inheritance of variation in future generations allow a change in the distribution of fitness values in subsequent generations. When any individual solution can be encoded in the form of a genotype/chromosome, the evolutionary process of selection, hybridization and recombination, can be used to create fitness values in a population when these processes are iterated over cycles of selection. The populations of solutions move towards global maximum to provide solutions to the optimization challenge.

The concept of an ideal genotype for any specific environment assumes the existence of one genotype configuration with the maximum fitness value for any one environment and the realization of this ideally adapted genotype is often compared to the OneMax problem in GA literature, where the population moves towards one global maximum. While this is considered an easy task when mutation is the only variation generating mechanism, the same OneMax problem could have many local optima if recombination among and within chromosomes provides a mechanism for generating genotypic variation. Fitness landscapes with many local optima are called multi-modal fitness landscapes and it is possible for populations be "trapped" at any of the local optima and never approach the global optimum (Wood et al. 2001; Du et al. 2015). Multi-modal functions are distinct from multi-objective optimization functions where several fitness functions are involved. To avoid the trap of local maxima, strategies like niching and sharing of

solutions have been proposed, which allow the population of solutions to escape from local optima and continue to search for the global optima (Goldberg 1989; Goldberg and Deb 1992; rey Horn 1997; Luque 2011).

There are several classes of GAs grouped based on the structure of the population and the distribution of connections among sub-populations. In canonical GA, all the individual solutions are pooled together in one population where crossing can occur with any individual solution and individual solutions for crossing are selected at random. In distributed GAs, the structure among subsets of individual solutions is maintained and each subset evolves in its sub-domain (island) solution space leading to solutions in different domains. In Cellular GA (CGA), the individual solutions are arranged in a grid, where the breeding pool for any individual solution is its immediate neighborhood. In distributed CGA, the islands and the individuals within the island are arranged in a grid and crossing or migration is possible only in the neighborhood of each island (Luque, 2011).

Factors such as number of islands, island size, migration rate, migration frequency, degree of connectivity among neighboring islands, the emigrant genotypes and the resident genotypes that are replaced in an island affect convergence time and quality of solutions of IM GA (Cantu-paz 2000; Whitley 1999; Skolicki 2007). Island Size & Number of islands: Larger island population size leads to better quality solutions, as it allows evolution of diverse solutions. It also shows a slower loss of genetic variance. But the tradeoff with convergence rate is linked with the number of islands. Inter-island diversity increases as the number of islands increases. This leads to an increase in global diversity of the population. But its effect on convergence time and solution quality is dependent on other factors such as degree of connectivity, migration size and frequency (Cantu-paz 2000; Whitley 1999; Skolicki 2007). Large numbers of migrants

replaces most of the selected solutions in the immigrant islands leading to a decrease in inter-island diversity, as copies of emigrant solutions spread rapidly. This leads to rapid convergence in most islands to sub-optimal solutions. Intermediate numbers of migrants perform the best in terms of both convergence rate and quality of solution. Quality of solutions can be assessed using any one of distance based metrics, non-dominance based metrics or spread based metrics (Goel and Stander 2010; Abouhawwash et al. 2020). The magnitude of intermediate size range depends on the island size and selection intensity (Cantu-paz 2000; Whitley 1999; Skolicki 2007). Optimal migration frequencies are intermediate between extremes. There is better diversity with longer migration intervals, as the local populations get time to evolve their own solutions. But very low frequencies of migration fails to rescue the local populations from stagnation (Cantu-paz 2000; Whitley 1999; Skolicki 2007).

Migration policies refer to decisions about individual emigrates and whether they are used to replace resident genotypes on in the island. When best solutions replace worst solutions on the immigrant island, the proportion of good solutions in the selected population increases by increasing the selection intensity in the immigrant islands. When worst solutions replace best solutions in the immigrant island, selection intensity decreases. When randomly selected emigrants replace random immigrants, there is no change in selection intensity and doesn't affect the convergence rate and quality of solutions and IM models perform similar to discrete selection where there is no migration (Cantu-paz 2000; Whitley 1999; Skolicki 2007).

While WGS and IM methods have been evaluated with truncation selection, where lines with criterion value above a threshold are selected, there are limits to truncation selection. Other selection methods similar to truncation selection such as i) tournament selection, where winner of k-size tournament is selected, ii) Ranking selection, where the probability of selecting

individuals depends on their ranking, iii) Fitness proportional, where the probability of selecting individuals depends on their fitness scores, and iv) Uniform method, where individuals are selected randomly (Cantu-paz 2000; Skolicki 2007) have been evaluated. While each of these selection methods have their advantages, these methods still have the limitations of truncation selection. In all these methods, relationships among selected genotypes are ignored and no constraints are imposed on inbreeding rate or genetic variance among progeny which is important for maintaining variance in the population for long-term genetic improvement (Brisbane & Gibson 1995). Selection methods based on multi-objective framework that takes into account multiple objectives and constraints to find solutions that are optimal trade-offs between conflicting objectives offer several advantages over truncation selection. In the third and fourth chapter of this dissertation, we have compared WGS on admixed populations and island populations.

Multi-objective Optimization Based Selection

Most of the initial studies that explored the application of constrained optimization and multi-objective optimization methods in the context of PS in breeding programs were done in animal systems. (Wray and Goddard 1993; Brisbane & Gibson 1995; Meuwissen 1997). With the availability of GS methods, several researchers have extended the exploration of these methods to GS in animal breeding (Daetwyler et al. 2007; Sonesson et al. 2010; Kinghorn 2011; Wooliams et al. 2015). The application of MOOB methods in plant breeding in the context of GS has been investigated by Akdemir et al and Gorjanc G et al (Akdemir and Sanchez 2015; Gorjanc et al. 2018; Allier et al. 2019d, 2019e). These three MOOB methods differ in the specification of multi-objective functions. While the genomic mating method minimizes inbreeding and co-ancestry and maintains a minimum expected genetic gain, the Optimal Cross

Selection (OCS) maximizes genetic gain while assuring a specified rate at which genetic variance is lost per cycle of recurrent selection (Akdemir and Sanchez 2015; Akdemir et al. 2018; Gorjanc et al. 2018). The genomic mating method resulted in greater long-term response relative to GS and PS for both single trait and multi-trait selection (Akdemir and Sanchez 2015; Akdemir D et al. 2018). Usefulness Criterion Parental Contribution based on OCS, UCPC-OCS, accounts for within-family variance and selection intensity while implementing the OCS strategy. UCPC-OCS method resulted in greater long-term responses relative to OCS method in multi-parental maize breeding programs (Allier et al. 2019d, 2019e).

Genetic Algorithms (GA) can be used to solve multi-objective optimization (MOGA) problems. In the case of MOGA, individual solutions are mapped to more than one fitness value and objective functions are constructed to account for the multiple objectives with constraints based on realistic breeding objectives. Solutions to MO problems fall on the pareto-optimal frontier which refers to a set of solutions that provide optimal tradeoff between the objectives. It effectively means the solutions on the pareto-frontier cannot be improved for any one objective without compromising on the other objective (Goldberg 1989; Goldberg and Deb 1992; rey Horn 1997; Luque 2011).

By setting objectives for maximization of genetic contributions, genetic gain, usefulness criteria and minimization of inbreeding in populations, it is possible to improve long-term response relative to truncation selection on any criteria. However, in most cases the percent gain in response with MOO approaches is relatively lower or equivalent to other methods of maintaining population genetic diversity (Akdemir and Sanchez 2015; Gorjanc et al. 2018). Moreover, GM and OCS methods have not been compared in a standard system for their performance.

1.5 Description of Dissertation Chapters

The projects described in this dissertation evaluated methods for prediction, selection of genotypes and selection of crosses to improve genetic gains in both the short term and over many cycles of simulated recurrent selection of soybeans adapted to maturity zones II and III in North America.

In Chapter 2, we examined the impact of five factors and their interactions on responses to 40 cycles of recurrent selection beginning with founders from the SoyNAM project (citation Diers et al). The five factors include PS and four GS methods, training sets, number of QTL, heritability and selection intensity. Response metrics included two standardized response measures. Response standardized to maximum possible genotypic value for a given genetic model (Rs) and Response standardized to change in genotypic variance (RsVar) that is similar to efficiency of converting genetic gain per unit loss of genetic variance (Liu et al. 2015; Gorjanc et al. 2018). Interactions among the five factors significantly affect the rate of response and limits to selection response. If GP models are not updated after initial training, bayesian parametric methods performed better than ridge-regression and machine learning methods in terms of prediction accuracy and limits of response to selection. If GP models are updated by re-training using current and prior cycles of genotypic and phenotypic data, then ridge regression based GS demonstrated greater prediction accuracies and better genetic gains in both the short term and across all 40 cycles. Relaxed selection intensities resulted in greater responses in later cycles of selection but at slightly lower rates of genetic gain in early cycles when compared to stringent selection intensities.

In chapter 3, we have compared the impact of breeding strategies on responses to 40 cycles of recurrent selection beginning with founders from the SoyNAM project. The breeding

strategies consisted of selection methods and crossing designs. The selection methods include PS, GS, and WGS conducted simultaneously across all families of RILs sampled from the admixed population created each cycle. In contrast, the families of RILs created every cycle were also treated as islands and selection methods were practiced on the members of each island. The crossing designs include the chain rule, high frequencies of the best selected lines, random mating and 'genomic mating'. In this study, we found that response pattern and limits of response are distinctive when we apply selection methods to the admixed population of RILs from application of selection methods to islands, i.e., individual families. Limits of responses were also influenced factors such as migration policy and migration frequency among islands. Migration direction and migration size didn't have any significant impact on limits to response from recurrent selection. We have also compared the potential of WGS on a global population and within islands with exchange of lines at regular intervals.

In chapter 4, we have provided a summary of the dissertation projects and discussed potential future work to extend the application of a few selection strategies to populations of real soybean genotype data. The subsequent chapters describe the dissertation projects.

References

Abouhawwash, Mohamed, Mohammed Jameel, and Kalyanmoy Deb. **"A Smooth Proximity Measure for Optimality in Multi-Objective Optimization Using Benson's Method."** *Computers & Operations Research* 117 (2020/05/01/ 2020): 104900.

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J: **Multi-objective optimized genomic breeding strategies for sustainable food improvement.** *Heredity* 2019, **122:**672-683.

Akdemir D, Sánchez JI: **Efficient Breeding by Genomic Mating.** *Frontiers in Genetics* 2016, **7**.

Allier A, Lehermeier C, Charcosset A, Moreau L, Teyssèdre S: **Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection.** *Frontiers in Genetics* 2019, **10**.

Allier A, Moreau L, Charcosset A, Teyssèdre S, Lehermeier C: **Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression.** *G3: Genes|Genomes|Genetics* 2019, **9:**1469.

Bailey-Serres J, Parker JE, Ainsworth EA, Oldroyd GED, Schroeder JI: **Genetic strategies for improving crop yields.** *Nature* 2019, **575:**109-118.

Barton NH, Etheridge AM, Véber A: **The infinitesimal model: Definition, derivation, and implications.** *Theoretical Population Biology* 2017, **118:**50-73.

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J: **Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.).** *Plant Science* 2016, **242:**23-36.

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H: **Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures.** *Genetics Selection Evolution* 2012, **44:**3.

Beavis W, Beavis W: **The power and deceit of QTL experiments: lessons from comparative QTL studies.** 1994.

Bernardo R: **Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years.** *Crop Science* 2008, **48:**1649.

Bernardo R: **Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations.** *Crop Science* 2014, **54:**2586.

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, et al: **Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress.(RESEARCH)(Author abstract).** 2015, **55:**154.

Bhatia VS, Singh P, Wani SP, Chauhan GS, Rao AVRK, Mishra AK, Srinivas K: **Analysis of potential yields and yield gaps of rainfed soybean in India using CROPGRO-Soybean model.(Report).** *Agricultural and Forest Meteorology* 2008, **148:**1252.

Bijma P: **Long-term genomic improvement – new challenges for population genetics.** *Journal of Animal Breeding and Genetics* 2012, **129:**1-2.

Blows M, Walsh B: **Spherical Cows Grazing in Flatland: Constraints to Selection and Adaptation.** In *Adaptation and Fitness in Animal Populations: Evolutionary and Breeding Perspectives on Genetic Resource Management.* Edited by van der Werf J, Graser H-U, Frankham R, Gondro C. Dordrecht: Springer Netherlands; 2009: 83-101

Brisbane J, Gibson J: **Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions.** *International Journal of Plant Breeding Research* 1995, **91:**421-431.

Byrum J, Beavis B, Davis C, Doonan G, Doubler T, Kaster V, Mowers R, Parry S: **Genetic Gain Performance Metric Accelerates Agricultural Productivity.** *Interfaces* 2017, **47:**442-453.

Cantú-Paz E: *Efficient and accurate parallel genetic algorithms / by Erick Cantú-Paz.* Boston, Mass.: Boston, Mass. : Kluwer Academic Publishers; 2000.

Cassman KG, Dobermann A, Walters DT, Yang H: **MEETING CEREAL DEMAND WHILE PROTECTING NATURAL RESOURCES AND IMPROVING ENVIRONMENTAL QUALITY.** *Annu Rev Environ Resour* 2003, **28:**315-358.

Collins LM, Dziak JJ, Kugler KC, Trail JB: **Factorial Experiments: Efficient Tools for Evaluation of Intervention Components: Efficient Tools for Evaluation of Intervention Components.** *American Journal of Preventive Medicine* 2014, **47:**498-504.

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G: **Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction.** vol. 65. pp. 311-336. Melbourne; 2014:311-336.

Cooper M, Podlich D: **TheE(NK) model: Extending theNK model to incorporate gene-by-environment interactions and epistasis for diploid genomes.** *Complexity* 2002, **7:**31-47.

Cooper M, Podlich D, Micallef K, Smith O, Jensen N, Chapman S, Kruger N: **Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops.** *Quantitative genetics, genomics and plant breeding'(Ed MS Kang) pp* 2002**:**143-166.

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, et al: **Genomic prediction in CIMMYT maize and wheat breeding programs.** *Heredity* 2014, **112:**48-60.

Crow J: **Sewall Wright's place in twentieth-century biology.** *J Hist Biol* 1990, **23:**57-89.

Crow JF, Kimura M: **An introduction to population genetics theory.** *An introduction to population genetics theory* 1970.

Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, Hickey JM: **Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking.** *Genetics* 2013, **193:**347-365.

DeWitt TJ, Scheiner SM: *Phenotypic plasticity: functional and conceptual approaches.* Oxford University Press; 2004.

Du Y, Ma Q, Aoki K, Sakamoto M, Furutani H, Zhang Y-a: *Hitting Time Analysis of OneMax Problem in Genetic Algorithm.* 2015.

Emily C, Rex B: **Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers.** *The Plant Genome* 2013, **6**.

Fehr W: *Principles of cultivar development: theory and technique.* Macmillian Publishing Company; 1991.

Felsenstein J: **The effect of linkage on directional selection.** *Genetics* 1965, **52:**349.

Flint-Garcia S: **Genetics and Consequences of Crop Domestication.** *Journal of agricultural and food chemistry* 2013, **61**.

Franco GA, Mark AN, William DB, Scott MP, Jean-Luc J: **Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats.** *The Plant Genome* 2011, **4:**132-144.

Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245-257.

Goel, Tushar, and Nielen Stander. **"A Study of the Convergence Characteristics of Multiobjective Evolutionary Algorithms."** In *13th Aiaa/Issmo Multidisciplinary Analysis Optimization Conference*. Multidisciplinary Analysis Optimization Conferences: American Institute of Aeronautics and Astronautics, 2010.

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS: **Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection.** *Genetics* 2017, **206:**1675.

Goldberg DE: **Sizing populations for serial and parallel genetic algorithms.** In *Proceedings of the 3rd international conference on genetic algorithms*. Morgan Kaufmann Publishers Inc.; 1989: 70-79.

Goldberg DE, Deb K: **Massive multimodality, deception, and genetic algorithms.** *Urbana* 1992, **51:**61801.

Goldman IL: **The Intellectual Legacy of the Illinois Long-Term Selection Experiment.** In *Plant Breeding Reviews.* John Wiley & Sons, Inc.; 2010: 61-78

Gorjanc G, Gaynor RC, Hickey JM: **Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection.** *Theoretical and Applied Genetics* 2018, **131:**1953-1966.

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G: **The impact of population structure on genomic prediction in stratified populations.** *Theoretical and Applied Genetics* 2014, **127:**749-762.

Guo Z, Tucker DM, Lu J, Kishore V, Gay G: **Evaluation of genome-wide selection efficiency in maize nested association mapping populations.** *Theoretical and Applied Genetics* 2012, **124:**261-275.

Guzman C, Peña RJ, Singh R, Autrique E, Dreisigacker S, Crossa J, Rutkoski J, Poland J, Battenfield S: **Wheat quality improvement at CIMMYT and the use of genomic selection on it.** *Applied & Translational Genomics* 2016, **11:**3-8.

Habier D, Fernando RL, Dekkers JCM: **The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values.** *Genetics* 2007, **177:**2389.

Hagan S, Knowles J, Kell DB: **Exploiting Genomic Knowledge in Optimising Molecular Breeding Programmes: Algorithms from Evolutionary Computing (Evolutionary Computing for Molecular Breeding).** 2012, **7:**e48862.

Harlan Wood D, Chen J, Antipov E, Lemieux B, Cedeño W: *A design for DNA computation of the OneMax problem.* 2001.

Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME: **Genomic selection accuracy for grain quality traits in biparental wheat populations.(RESEARCH)(Author abstract)(Report).** *Crop Science* 2011, **51:**2597.

Heslot N, Jannink J-L, Sorrells ME: **Perspectives for Genomic Selection Applications and Research in Plants.** *Crop Science* 2015, **55:**1-12.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Science* 2012, **52:**146-160.

Hickey JM, Bruce C, Whitelaw A, Gorjanc G: **Promotion of alleles by genome editing in livestock breeding programmes.** *Journal of Animal Breeding & Genetics* 2016, **133:**83-84.

Hickey JM, Chiurugwi T, Mackay I, Powell W: **Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery.** *Nature genetics* 2017, **49:**1297.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G: **Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation.(RESEARCH)(Author abstract).** 2014, **54:**1476.

Howard R, Carriquiry AL, Beavis WD: **Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures.** *G3 (Bethesda)* 2014, **4:**1027-1046.

Jannink J-L: **Dynamics of long-term genomic selection.** *Genetics Selection Evolution* 2010, **42:**35.

Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, Hickey JM: **Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs.** *Genetics Selection Evolution* 2015, **47:**55.

Jonas E, de Koning D-J: **Does genomic selection have a future in plant breeding?** *Trends in Biotechnology* 2013, **31:**497-504.

Jonas E, de Koning DJ: **Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops.** *Biotechnology & genetic engineering reviews* 2016, **32:**18.

Kempthorne O: ***An introduction to genetic statistics***. Iowa State University Press; 1969.

Keurentjes JJB, Molenaar J, Zwaan BJ: **Predictive modelling of complex agronomic and biological systems.** *Plant, Cell & Environment* 2013, **36:**1700-1710.

Kinghorn BP: **An algorithm for efficient constrained mate selection.** *Genetics Selection Evolution* 2011, **43:**4.

Liu H, Meuwissen TH, Sorensen AC, Berg P: **Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs.** *Genet Sel Evol* 2015, **47:**19.

Liu Z, Yang X, Lin X, Hubbard K, Lv S, Wang J: **Maize yield gaps caused by non-controllable, agronomic, and socioeconomic factors in a changing climate of Northeast China.** *Science of The Total Environment* 2016, **541:**756-764.

Lobell DB, Cassman KG, Field CB: **Crop yield gaps: their importance, magnitudes, and causes.** *Annual review of environment and resources* 2009, **34:**179-204.

Logan A: **Will Agricultural Technofixes Feed the World? Long- and Short-Term Tradeoffs of Adopting High-Yielding Crops.** In; 2017

Long N, Gianola D, Rosa GJM, Weigel KA: **Application of support vector regression to genome-assisted prediction of quantitative traits.** *Theoretical and Applied Genetics* 2011, **123:**1065.

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, GonzÁLez-Recio O: **Radial basis function regression methods for predicting quantitative traits using SNP markers.** *Genetics Research* 2010, **92:**209-225.

Luque G: **Parallel Genetic Algorithms: Theory and Real World Applications.** Berlin, Heidelberg : Springer Berlin Heidelberg; 2011.

Marulanda J, Mi X, Melchinger A, Xu J-L, Würschum T, Longin C: **Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale.** *Theor Appl Genet* 2016, **129:**1901-1913.

Mathan J, Bhattacharya J, Ranjan A: **Enhancing crop yield by optimizing plant developmental features.** *Development* 2016, **143:**3283.

Mathieu C, Bernard A: **Wright's shifting balance theory and the diversification of aposematic signals.** *PLoS ONE* 2012, **7:**e34028.

Meuwissen T, Hayes B, Goddard M: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819-1829.

Meuwissen TH: **Maximizing the response of selection with a predefined rate of inbreeding.** *Journal of animal science* 1997, **75:**934-940.

Moose SP, Dudley JW, Rocheford TR: **Maize selection passes the century mark: a unique resource for 21st century genomics.** *Trends Plant Sci* 2004, **9:**358-364.

Morrissey M B, Kruuk L. E B, Wilson A J: **The danger of applying the breeder's equation in observational studies of natural populations.** *Journal of Evolutionary Biology* 2010, **23:**2277-2288.

Myers RH: *Response surface methodology / Raymond H. Myers.* s.l. : [Ann Arbor: s.l. : s.n.

Ann Arbor : available from Edwards Brothers; 1976.

Nakaya A, Isobe SN: **Will genomic selection be a practical method for plant breeding?** *Annals of Botany* 2012, **110:**1303-1316.

Okasha S: **Fisher's Fundamental Theorem of Natural Selection—A Philosophical Analysis.** *The British Journal for the Philosophy of Science* 2008, **59:**319-351.

Pigliucci M: **Sewall Wright's adaptive landscapes: 1932 vs. 1988.** *Biol Philos* 2008, **23:**591-603.

Pigliucci M: **'On the Different Ways of ''Doing Theory'' in Biology'.** *Biological Theory* 2013, **7:**287-297.

Pitchers W, Wolf JB, Tregenza T, Hunt J, Dworkin I: **Evolutionary rates for multivariate traits: the role of selection and genetic variation.** *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2014, **369:**20130252.

Podlich DW, Cooper M: **Modelling Plant Breeding Programs as Search Strategies on a Complex Response Surface.** In *Simulated Evolution and Learning: Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL'98 Canberra, Australia, November 24–27, 1998 Selected Papers.* Edited by McKay B, Yao X, Newton CS, Kim J-H, Furuhashi T. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999: 171-178

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ: **Empirical fitness landscapes reveal accessible evolutionary paths.** *Nature* 2007, **445:**383.

Price GR: **The nature of selection.** *Journal of Theoretical Biology* 1995, **175:**389-396.

rey Horn J: **The Nature of Niching:, Genetic Algorithms and the Evolution of Optimal, Cooperative Populations.** *Urbana* 1997, **51:**61801-62996.

Rutkoski JE: **Chapter Four - A practical guide to genetic gain.** In *Advances in Agronomy. Volume* 157. Edited by Sparks DL: Academic Press; 2019: 217-249

Sheftel H, Shoval O, Mayo A, Alon U: **The geometry of the Pareto front in biological phenotype space.** *Ecology and Evolution* 2013, **3:**1471-1483.

Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U: **Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space.** *Science (New York, NY)* 2012, **336:**1157.

Skolicki Z, Jong KD: **The importance of a two-level perspective for island model design.** In *2007 IEEE Congress on Evolutionary Computation*; *25-28 Sept. 2007*. 2007: 4623-4630.

Skolicki ZM: **An analysis of island models in evolutionary computation.** George Mason University, 2007.

Smith JM: **Optimization Theory in Evolution.** *Annu Rev Ecol Syst* 1978, **9:**31-56.

Sonesson A, Woolliams J, Meuwissen T: **Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection.** In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 1*. 2010

van Ittersum MK, Cassman KG, Grassini P, Wolf J, Tittonell P, Hochman Z: **Yield gap analysis with local to global relevance—A review.** *Field Crops Research* 2013, **143:**4-17.

Wang J, van Ginkel M, Podlich D, Ye G, Trethowan R, Pfeiffer W, DeLacy IH, Cooper M, Rajaram S: **Comparison of two breeding strategies by computer simulation.** *Crop Science* 2003, **43:**1764+.

Whitley D, Rana S, Heckendorn RB: **The island model genetic algorithm: On separability, population size and convergence.** *CIT Journal of computing and information technology* 1999, **7:**33-47.

Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C: **Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection.** *Genetics* 2013, **195:**573.

Woolliams JA, Berg P, Dagnachew BS, Meuwissen TH: **Genetic contributions and their optimization.** *J Anim Breed Genet* 2015, **132:**89-99.

Wright S: **"Surfaces" of selective value.** *Proceedings of the National Academy of Sciences of the United States of America* 1967, **58:**165.

Wright S: **Surfaces of Selective Value Revisited.** *The American Naturalist* 1988, **131:**115-123.

Yabe S, Yamasaki M, Ebana K, Hayashi T, Iwata H: **Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops.** *PLoS One* 2016, **11:**e0153945.

CHAPTER 2. MODELING SIMULATED RESPONSES TO RECURRENT GENOMIC SELECTION IN SOYBEANS

Vishnu Ramasubramanian[*] [†] **and William D Beavis**[*†]

[*] Department of Agronomy

[†]Bioinformatics and Computational Biology Graduate Program

Iowa State University, Ames, Iowa, USA – 50010

Modified from a manuscript that has been conditionally accepted by Frontiers in Genetics

Abstract

What range of responses across multiple cycles of genomic selection (GS) should public soybean breeders expect? We applied a recurrence equation to model responses from a factorial set of selection methods, training sets, selection intensities, genetic architectures and heritabilities across 40 cycles of simulated recurrent selection of soybean lines derived from the Soybean Nested Association Mapping (SoyNAM) population. The model enabled detection of statistical significance due to selection methods, training sets, and selection intensities, which are under the control of the breeder, as well as genetic architecture and heritability, which are not. The estimated parameters of the model also enabled estimates of response rates, population half-lives and genotypic values as responses to recurrent selection approach asymptotic limits. The modeled responses revealed that both the rates of genetic improvement in the early cycles and limits to genetic improvement in the later cycles are significantly affected by interactions among all five factors. Even though all possible interactions significantly affected modeled responses, there were some consistent trends. In early cycles all genomic selection (GS) methods resulted in greater response rates than phenotypic selection (PS). Phenotypic selection produced greater genotypic values then GS as asymptotic limits are approached in later cycles. Updating with

training sets consisting of data from prior cycles of selection significantly improved prediction accuracy and genetic response for all GS methods. From among the GS methods with updated training sets, selection on estimates from Ridge Regression- Restricted Maximum Likelihood Estimation method achieved better response rates and longer half-lives than selection on estimates from BayesB and Bayes LASSO methods. A Support Vector Machine with a radial basis kernel method, resulted in the fastest loss of genetic variance in the early cycles. Before applying GS to soybean breeding populations the breeder should first explicitly determine the objectives of the project because achieving maximum short term responses require different combinations of selection methods, training sets, and selection intensities than combinations that produce maximum responses at the limits of response to selection.

1. Introduction

Soybeans has been subjected to recurrent phenotypic selection for several thousand years (Anderson et al. 2019). Current genetic improvement in commercial and public soybean breeding integrates variety development as the evaluation phase of recurrent selection (Orf 2008). Each cycle of selection consists of making crosses among selected lines, creating replicable homozygous lines, evaluating the lines in replicated field trials during which poor performing lines are discarded, and crossing the retained lines to begin a new cycle. Pedigrees of modern soybean varieties confirm that genetic improvement of soybeans in Maturity Zones (MZs) II, III and IV has been largely through intra-population recurrent selection (Hyten et al. 2006; Mikel et al. 2010; Langewisch et al. 2017; Achard et al. 2020). Prior to the Plant Variety Protection Act in 1970 (https://www.ams.usda.gov/rules-regulations/pvpa), a cycle of genetic improvement for soybeans would require 12 to 14 years, whereas in the last 40 years commercial organizations have invested in development of continuous nurseries including software that streamlines

inventories and logistics of seed transfer, resulting in the capacity to complete a cycle of recurrent selection in five years (Byrum et al. 2017; Anderson et al 2019). Thus, using current practices a soybean breeder might experience three to five cycles of genetic improvement.

Despite reducing cycle times to less than ½ of the time required 40 years ago, yield improvements for soybean have not doubled in the corn-soybean agricultural production region of the United States (Mikel et al. 2010; Specht et al. 2014). To address the relatively slow genetic improvements in soybean Genomic Selection (GS) was implemented in commercial soybean variety development projects (Kurek, 2018). The expectation is to increase the rate of genetic improvement by increasing selection intensity by enabling evaluation of larger numbers of progeny without the commensurate added expense of increased numbers of field plots (Heslot et al., 2012). However, it is well-established that increased selection intensities will reduce the genetic potential of a breeding population through loss of useful genetic variability and ultimately will limit responses to selection (Robertson 1960; Hill and Robertson 1968; Bulmer 1970). Recognizing the potential impact on loss of useful genetic diversity from application of GS to elite soybean varieties, farmer-members of the North Central Soybean Research Program (NCSRP) provided funding to public soybean breeders to identify novel QTL in public genotypes adapted to MZ's II, III and IV (Song et al. 2017). The project provided evidence of novel useful genetic variability for yield, but favorable alleles were widely scattered among the SoyNAM lines (Diers et al. 2018). Subsequently, NCSRP provided funding to use GS to improve public germplasm and reduce the gap between public sources of novel favorable alleles and elite commercial varieties (McHale et al. 2017). An obvious question to ask is "What range of responses across multiple cycles of GS should public soybean breeders expect from the SoyNAM founders"?

After the seminal work by Bernardo and Yu (2007), crop breeders have experimentally demonstrated the relative advantage of GS to phenotypic selection (PS) for initial cycles of recurrent selection in barley, maize, oats, rice and wheat (Bernardo 2008, 2014; Asoro et al. 2011; Heslot et al. 2012, 2015; Nakaya and Isobe 2012; Emily and Bernardo 2013; Crossa et al. 2014; Beyene et al. 2015; Bassi et al. 2016; Marulanda et al. 2016; Jonas and de Koning 2013, 2016; Hickey et al. 2017; Goiffon et al. 2017). While these results demonstrate rapid responses in the first few cycles of genetic improvement, they did not address the loss of genetic potential from the founder populations, nor how the observed responses could have been affected by the GS method, population structure, genome organization, possible training sets, etc.

Simulations have been used to demonstrate that genome organization, population structure, genetic architecture, heritability, selection intensity, Genomic Prediction (GP) models and composition of training sets will affect responses to GS (Habier et al. 2007; Goddard 2009; Zhong et al. 2009; Jannink 2010; Bastiaansen et al. 2012; Bijma 2012; de los Campos et al. 2010; 2013; Wimmer et al. 2013; Howard et al. 2014; Liu et al. 2015; Michel et al. 2016). Most studies have evaluated impacts of subsets of factors on responses across a few cycles of recurrent selection, although a couple of studies investigated limits of responses to Recurrent GS (RGS) across more than 20 cycles (Jannink 2010; Liu et al. 2015). Linear mixed model (LMM) methods provide more accurate predictions and greater gains than non-parametric machine-learning methods for traits with additive genetic architecture. Prediction accuracies are similar for all GP models applied to simulated additive genetic architectures among numbers of genotypes typically used by plant breeders (Long et al. 2010, 2011; Guo et al. 2012; Howard et al., 2014) although Bayesian methods provide greater responses after about 15 cycles of RGS (Pérez and de los Campos, 2014).

Simulations also have revealed that marker densities impact prediction accuracies and short term genetic gains, with a dense set of markers performing better than a sparse set. There are limits to improvements from increased marker densities that depend on linkage disequilibrium (LD) and structure of the breeding population (de Roos et al. 2009; Schulz-Streeck et al. 2012; Hickey et al. 2014; Xavier et al, 2016; Norman et al. 2018). Decreased prediction accuracies of GP models in later cycles of RGS are associated with decay of LD between marker loci (ML) and quantitative trait loci (QTL), loss of relationships between lines in early and later cycles of selection or a combination of both (Habier et al. 2007; Zhong et al. 2009; Hickey et al. 2014; Liu et al. 2015; Müller and Melchinger 2017, 2018). It is possible to offset the loss of relationships between training and validation sets as well as loss of LD, by updating the training sets with each cycle of progeny from the selected lines (Jannink 2010; Liu et al. 2015; Müller et al. 2017, 2018). If training data are from only the current cycle of selection, then predictions do not take into account relationships between the current population and the founder population or possible loss of LD. At the other extreme, data from all prior cycles of selection can be included with data from the current cycle in the training set, but computer memory limits can be encountered with large training sets.

Curiously, responses of RGS affected by the various factors have not been modelled. Because responses to RGS and PS are non-linear, non-decreasing functions of discrete, i.e., integer values, of cycles, where the responses for each cycle after the founding generation depend on values of the prior cycle, recurrence equations, a.k.a. difference equations (Goldberg 1958), should be ideal for modeling. Assuming that a recurrence equation provides a good fit to the response values, the model should provide the ability to assign statistical significance to hypotheses about contributions of factors to variability among responses, and the model's

parameters can be used to provide estimates of response rates at any point in time, estimates of population half-lives and estimates of asymptotic limits to responses.

While current practices used in soybean breeding might require as much as five years per cycle of selection, the development of genetic male sterility and double haploid technologies for soybean will make it possible to conduct a cycle of RGS every year (Ortiz-Perez 2008; Davis and Byrum, 2014). In anticipation of emerging technologies that will enable public soybean breeders to implement GS with ever shorter cycle times, we used simulations to investigate the impacts of selection methods (PS and GS), training sets (TS), number of QTL (nQTL), heritability (H), selection intensity (SI) and their interactions on responses to RGS applied to a small genetic improvement program founded on 2000 lines derived from crosses involving 20 SoyNAM founders. We employed a factorial treatment design, where each combination of factor levels was replicated ten times in simulated RGS and RPS across 40 cycles and modeled the responses using a first order recurrence equation. The modeled responses address the question about the range of responses that a public soybean breeder can expect from GS applied to lines derived from SoyNAM. Also the range of modeled responses provide a foundation for comparing soybean GS strategies and practical guidelines for understanding trade-offs between genetic gains in early cycles and retention of useful genetic potential for future genetic improvement of soybeans.

2. Methods

2.1 Simulations and Treatment Design.

The impact of nQTL, SI, H, TS and SM on response to selection across 40 cycles of recurrent selection were evaluated with a factorial design consisting of 306 combinations of factors. In contrast to evaluating one or two variables at a time, the factorial design is widely used to comprehensively determine factors that will optimize processes (Myers 1976; Collins et

al. 2014; Dunn 2020). Specifically the treatments consisted of three values for number of simulated QTL, three selection intensities, two values for non-genetic variance, five selection methods and four types of TS used to update four GP models (Table 1). In summary the treatment design consists of 18 combinations of factors for PS plus 288 combinations of factors for GS methods for a total of 306 combinations of factors. Note that TS are irrelevant for PS. Each set of factor combinations was replicated with ten simulated recurrent selection projects across 40 cycles resulting in 3060 simulations with 122400 outcomes.

Simulated soybean lines were generated by crossing *in silico* 20 homozygous SoyNAM founder lines with IA3023 to generate 20 distinct F_1 progeny. The F_1 progeny from each of the 20 crosses were self-pollinated *in silico* for five generations to generate 100 lines per family. The resulting 2000 lines from 20 families were assessed for segregating genotypic information at 4289 SNP loci (Song et al. 2017; Xavier A et al. 2017; Diers B et al. 2018). On average alleles from the common founder occurred at a frequency of 0.9 and alleles at the loci from all other founder lines occurred at a frequency of 0.1.

Of the 4289 SNP markers with genotypic scores for the SoyNAM population 3818 were polymorphic among the 20 families used as founders for the simulations. On average, 773 were polymorphic for a family with a standard deviation among families of 34. In the initial founding set of RILs, the average heterozygosity per SNP locus across 20 families was 0.09. 'Gst' is a measure of sub-population differentiation estimated as ratio of difference between expected heterozygosity of sub- populations to total expected heterozygosity. The average estimated G_{st} value across the genome for the initial founding set of RILs was 0.32, as determined by the 'diff_stats' function in the mmod R package (Jombart 2008; Ryman and Leimar, 2009; Jombart and Ahmed, 2011). Relative to previous simulation studies that used a coalescent process

(Woolliams and Corbin, 2012; Hickey and Gorjanc, 2012; Daetwyler et al., 2013), our simulations began with less, albeit more realistic, genetic diversity for soybeans adapted to maturity zones II, III and IV. We also estimated Pairwise 'Fst' using 'pairwise.fst' in 'hierfstat' R package (Goudet 2005), which is a measure of population differentiation among pairs of populations. It is estimated as the ratio of difference between the average of the expected heterozygosity of the two populations and total expected heterozygosity of the pooled populations to total expected heterozygosity of the pooled populations. Average Fst among the 20 families in simulated SoyNAM data is 0.20. Whereas the average Fst using genotypic data from SoyNAM project among 40 families is 0.09 with a maximum pairwise Fst of 0.15 and a minimum Fst of 0.007. Average Fst among the clusters in USDA soybean germplasm collection is 0.22 -0.23 (Song et al 2015; Xavier 2018).

Subsets of 40, 400, and 4289 SNP loci were designated as QTL. The QTL were distributed evenly throughout the genome, and each contributed equal additive effects of 5/-5, 0.5/-0.5, or 0.05/-0.05 units respectively to the total genotypic value of the simulated RILs. Thus, all three genetic architectures had the same potential to create genotypic values ranging from +200 to -200 in the initial founder sets of RILs. Positive and negative allelic effects were simulated to alternate sequentially at QTL that are uniformly distributed across the Soybean genetic map. Because all marker alleles are QTL alleles when there are 4289 QTL, LD between marker alleles and QTL will not deteriorate across cycles of selection and recombination. Phenotypic values were simulated by adding non-genetic variance sampled from an N (0, σ) distribution to the simulated genotypic values, where σ was determined by the heritability on an entry mean basis among the initial sets of founder sets of RILs. Broad sense heritability on an entry mean basis (H) values of 0.7 and 0.3 were simulated for each of the three sets of QTL.

After the phenotypic values were simulated in the initial founding RILs, the non-genetic variance was held constant across subsequent cycles of selection.

For each cycle of recurrent selection, 1%, 2.5% or 10% of the most positive phenotypic or predicted phenotypic values among 2000 simulated RILs, corresponding to selection intensities of 2.67, 2.34 and 1.75, were selected to create lines for the subsequent cycle. Top ranked lines with the greatest phenotypic or GBLUP values are mated the most to generate the next cycle of progeny (Figure 1), referred herein as a hub network mating design (Guo et al. 2013; Guo et al. 2014). Population size was kept constant at 2000 lines for every cycle. This is a small number compared to commercial genetic improvement projects, but a reasonable number for resource limited public breeding projects.

Based on previous results from Howard et al (2014), four GS methods were evaluated. Ridge Regression using restricted maximum likelihood, represented a frequentist parametric model. Bayes-B (BB) and Bayesian LASSO (BL) represented parametric bayesian models and Support Vector Machine with Radial Basis Kernel (SVM-RBF) represented a non-parametric method of machine learning. Ridge regression was implemented with a method that employs expectation maximization to obtain Restricted Maximum Likelihood estimates of marker effects (Xavier et al. 2019). This computational method is faster than the popular implementation of ridge regression in the rrBLUP package (Endelman 2011) and generates values that are highly correlated with RRBLUP values from the rrBLUP package (Supplementary Figure 1). The BGLR package (Perez and de los Campos 2014) provided implementations of BB and BL models. The ‘Rgtsvm’ package in R was used for its implementation of the SVM with RBF kernel method (Wang et al. 2017). ‘Rgtsvm’ implements SVM training on GPUs with computing time several hundred times less than that required for the implementation in the ‘caret’ package

on high performance computing clusters, and produces similar prediction accuracies and estimates of mean squared errors (Supplementary Figure 2). The parameters used to train GP models with R packages are provided in Table 2.

For purposes of this manuscript we use the phrase 'model updating' to refer to retraining GP models with up to 14 previous cycles of training data (Supplementary Figure 3). A preliminary analysis of TS on genotypic values and prediction accuracies was conducted using RR-REML models trained with data from the current cycle as well as 3, 5, 8, 10, 12, and 14 prior cycles. The results were compared with responses from the RR-REML model updated with TS's comprised of data from all prior cycles and a RR-REML model with no updating. Training sets for each cycle were obtained by randomly sampling 1600 RILs from the set of 2000 simulated RILs in each cycle. The most accurate prediction and maximum genetic response was obtained with training data that is cumulatively added every cycle (Supplementary Figure 4 and S5). The results indicate that 3-5 prior cycles of training data did not significantly improve prediction accuracies and responses relative to models that were not updated. Also, the standardized genotypic values and prediction accuracies, obtained using 10 to 14 prior cycles of data in the TS's, were not significantly different than results based on TS's consisting of all prior cycles. Based on the results of this preliminary study, we investigated responses to recurrent selection using TS's consisting of up to 14 prior cycles of selection as well as data from the current cycle. After the 14th cycle, training data consisted of 14 prior cycles of recurrent selection.

2.2 Modeled Response to Recurrent Selection

The averaged genotypic value for each cycle, c, of recurrent selection was modeled with a linear first order recurrence equation:

$$f_0(\mathrm{c})y_{(\mathrm{c}+1)} + f_1(\mathrm{c})y_{(\mathrm{c})} = g(\mathrm{c}) \qquad \text{(Eqn 1)}$$

Where c is a sequence of integers from 0 to 39 representing each cycle of recurrent selection from cycle 1 to 40 and f_0, f_1 and g are constant functions of c. By rearranging the equation we note that the response in cycle c+1 can be represented as

$$y_{(c+1)} = -\frac{f_1(c)}{f_0(c)} y_{(c)} + \frac{g(c)}{f_0(c)} \qquad \text{(Eqn 2)}$$

Since the ratios $f_1(c)/f_0(c)$ and $g(c)/f_0(c)$ are constants, we can represent the response in cycle c+1 as

$$y_{(c+1)} = \alpha y_{(c)} + \beta \qquad \text{(Eqn 3)}$$

If y_0 specifies the average genotypic value of the first cycle of RILs derived from the founders, then (3) has a unique solution (Goldberg 1958):

$$y_c = \alpha^c y_0 + \beta \frac{1-\alpha^c}{1-\alpha} \text{ if } \alpha \neq 1 \qquad \text{(Eqn 4)}$$
$$y_c = \alpha^c y_0 + \beta c \quad \text{if } \alpha = 1$$

An alternative representation of (eqn 4) for the situation of $\alpha \neq 1$ is

$$y_c = \alpha^c (y_0 - y') + y'$$
$$with\ y' = \frac{\beta}{1-\alpha},$$

, where α is less than 1 for genotypic response to recurrent selection and y' represents the asymptotic limit to selection (Goldberg 1958). To illustrate, values of the sequence of c=0 to 39 for the range of α (0.6- 0.9) and β (1.4-38) values are plotted in Figure 2. The model derived curves can be interpreted as response to selection as a function of the frequencies of alleles with additive selective advantage, selection intensity, time and effective population size (Robertson 1960). The parameters α and β were estimated with a non-linear mixed effects method implemented in ‘nlme’ functions in the ‘nlme’ and ‘nlshelper’ packages (Pinheiro and Bates 2000; Baty et al. 2015; Pinheiro et al. 2019).

Since the limits of responses are approached asymptotically, the number of cycles required to reach half of the limits before there is no longer response to selection is referred to as the half-life of the recurrent selection process is referred to as the half-life (Robertson 1960; Dempfle 1974; Kang 1979; Cockerham & Burrows 1980; Kang and Namkoong 1980; Kang 1987; Kang and Nienstaedt 1987). From the first order recurrence equation, the half-life is estimated as ln (0.5)/ln (α), when y0 is '0' and the asymptotic limit is estimated as y'. When $y_0 \neq 0$, half-life is estimated as ln (0.5 * (y'/y'-y_0) / ln (α). The assumptions underlying the use of recurrence equations and correspondence with the theory of limits described by Robertson is provided in Supplementary File 1.

2.3 Analyses of variance (ANOVA) of Modeled Response to Recurrent Selection

The purpose of the ANOVA is to evaluate the impact of factors and their interactions on the modeled responses to recurrent PS and GS methods. The analyses of variance used single and multi-level nlme models with modeled (eqn 4) responses grouped by treatment factors. The influence of multiple factor treatment combinations on estimated non-linear mixed effect models have not been implemented in standard statistical software packages that report the analysis of variance in terms of sums of squares and traditional 'F-tests'. For discussions on the challenges of using standard F-test for non-linear mixed effects (nlme) models see (Pinheiro et al. 2000; Baty et al. 2015; Pinheiro et al. 2019). Consequently, we analyzed the variance among modeled responses using AIC, BIC and Likelihood metrics that were grouped based on combinations of factors consisting of selection methods, TS, SI, nQTL and simulated H.

In order to provide a balanced data table for analyses by the non-linear mixed effect model, responses that included PS, which has no TS's, were assumed constant resulting in a balanced full factorial set of responses for 360 combinations of factors. The process of fitting,

selecting and refining mixed effects models closely followed the steps described in Pinheiro et al. 2000; Zuur 2009 and Oddi et al. 2019). The complete description of the process used to fit NLME models and perform ANOVA is provided in Supplementary File 2.

In the first phase of model fitting, estimates of modeled parameters from nlsList models were retained as starting values for fixed effects. Both alpha and beta were fit only for intercept and deviations from estimated means conditioned on grouping variables were modeled as random effects using the 'nlme' R package. Multiple ANOVA of 'nlme' objects representing the models were used to identify combinations of factors with significant effects on the non-linear response model. The model with the lowest AIC score was selected as the best model. The best random intercept model in the first phase of model fitting process M31 in Supplementary Table 1 was further refined by modelling the correlation structure.

2.4 Evaluation of Simulated Response to Recurrent Selection

While the modeled genotypic values are evaluated using half-life and asymptotic limits, we have evaluated the simulated outcomes from recurrent selection using a set of metrics to assess responses, population characteristics, and GP model performance every cycle of selection.

The standardized genotypic value, R_s (eqn 5), was estimated every cycle as the change in genotypic value from the average genotypic value of 2000 RILs derived from the initial founders and standardized to the maximum genotypic potential (200 units) among the founders (Meuwissen et al. 2001; Liu et al. 2015).

$$R_s = \frac{R_c}{(R_m - R_0)} \qquad \text{(Eqn 5)}$$

R_s - Standardized genotypic value
R_0 - Average genotypic value of RILs produced by founders
R_c - Average genotypic value in cycle c – R_0
R_m - Maximum possible genotypic value (=200)

The most positive genotypic value (M_{gv}) among RILs selected in cycle c is a metric used to evaluate the best RIL produced each cycle, while the standardized genotypic variance (*Sgv*) defined as the estimated genotypic variance divided by the estimated genotypic variance of the initial population, was used to evaluate the loss of genotypic variance. Note that values for the *Sgv* range from zero to one.

Response standardized to change in standard deviation of genotypic values captures genetic gain with respect to loss of genetic variance (eqn 6). The numerator term represents change in genotypic values of a population in cycle 'c' from cycle '0' founder population normalized to standard deviation of genotypic values in cycle '0'. The denominator term represents change in standard deviation of genotypic values from cycle '0' to cycle 'c' as a fraction of standard deviation of genotypic values in cycle 0. This metric is similar to the metric used to refer to efficiency of converting loss of genetic diversity to genetic gain of a selection method in recurrent selection (Gorjanc et al. 2018). While efficiency is estimated as slope in linear regression model with numerator as 'y' term and denominator as 'x' term in the linear part of response curve, with Rs_Var it is possible to visualize both linear and non-linear sections of the response curve.

$$\text{Rs\_var} = \frac{G_c - G_0}{SdG_0 - SdG_c} \qquad \text{(Eqn 6)}$$

G_c-average genotypic value of the set of RILs evaluated in cycle c
G_0-average genotypic value of the founding set of RILs
SdG_0 - estimated standard deviation from genotypic values of founding set of RILs
SdG_c - estimated standard deviation from genotypic values of RILS in cycle c

Estimated Linkage disequilibrium (*LD*) among pairs of marker loci on all 20 chromosomes was evaluated as the deviation of observed gametic frequency of alleles at a pair of loci from the product of the individual allele frequencies, assuming independence (Weir 1996). The R function 'get_PG_LD_StatsGSMethods' used to estimate pairwise LD between markers is provided in the R package 'SoyNAMPredictionMethods'. GP models were assessed using the estimated prediction accuracies (r_{ps}), defined as the estimated linear correlation (Pearson) between predicted and simulated phenotypic values and the estimated Mean Squared Error (MSE), and defined as the mean of the squared deviations of the predicted phenotypic values from the simulated values.

3. Analyses and Data Availability

More information on the analyses can be found in the R package 'SoyNAMPredictionMethods'. Simulated data and code are also available as part of the package. Supplemental material including the R package can be found at http://gfspopgen.agron.iastate.edu/SoyNAMPredictionMethods_v2_2020.html. A complete description of the process for fitting NLME models and ANOVA can be found in Supplementary File 2. SoyNAM genotypic and phenotypic data are available in SoyBase (Grant et al., 2010).

4. Results

4.1 Prediction Accuracies in the Founding Sets of RILs

Estimates of prediction accuracies, r_{ps}, of GP models trained with the initial set of 2000 F_5-derived RILs ranged from 0.75-0.82 for H of 0.7 and ranged from 0.38 - 0.49 for H of 0.3 (Figure 3). The initial r_{ps} for both H values was best with BB and poorest with the SVM-RBF. The nQTL had little effect on r_{ps} within either value of 0.7 or 0.3 for H. RR-REML and BL produced smaller magnitude MSE values than BB and SVM RBF for all numbers of simulated QTL and both values for H (Figure 3). Accuracies are lower when GP models are trained without

markers assigned as QTL in the training set, but follow a similar pattern as the models that are trained with markers that are QTL. Estimates of MSE are greater or comparable for models trained without QTL than for models with QTL (Supplementary Figure 6- 7). Average within family prediction accuracies are less than prediction accuracies from a combined TS comprised of RILs from all the families (Supplementary Figure 8-9). However, a combined TS (n*population size of family) for 'n' families and estimated accuracies will have confounding effects from training set size. Estimated accuracies for models trained with TS generated by random sampling from a combined population to keep the TS size same as family size are lower than average within family accuracies. MSE are less for combined TS than for models trained using within family TS and sampled TS (Supplementary Figure 8-9).

4.2 Influence of Factors on Response Metrics

The averaged genotypic values were modeled (eqn 4) and the results are consistent with theory (Figure 2). Averaged across all simulations there was rapid increase of genotypic values across the first five cycles of selection followed by slower responses from cycles 5 to 10 and no response after cycle 20. While there are observable general trends for each of the individual factors, response metrics are unique for each combination of all factors (Supplementary Figure 10 -14). The most parsimonious model requires unique estimates of α, and β (eqn 4) for each of the combinations of factors indicating that interactions among all factors have significant influences on the responses (Supplementary Table 1). Also analyses of variance on subsets of 10 and 20 cycles of selection demonstrate that the interactions were important from the earliest cycles (Supplementary File 3: Supplementary Table 1 & 2). Further, the relative importance of factors on interaction effects were consistent (nQTL>SM>SI>H>TS) and significant in analyses of variance conducted on subsets of 10 and 20 selection cycles (Supplementary Table 1).

Among the three factors that are under the control of the plant breeder, we consider SM as the primary factor of interest, whereas TS and SI are considered secondary factors of interest that significantly modulate the effect of SM on response. In addition, nQTL and H, not under the control of the plant breeder, are noted for their large and significant impacts on response metrics (Supplementary Table 1).

4.2.1 The modeled responses were highly dependent on each of the factors, with nQTL showing the greatest deviations from the average response (Figure 2, Supplementary Table 1). From the modeled responses the half-life for a responsive population varied from 1.6 to 4.9 cycles for 40 QTL, whereas for 400 and 4289 QTL it ranged from 2.2 to 8.8 and 3 to 9.5 cycles respectively. The asymptotic limits ranged from 62.85 to 159.31 for 40 simulated QTL with a mean of 113.63, whereas for 400 simulated QTL the asymptotic limits ranged from 23.58 to 84.17 with a mean value of 50.89. For 4289 QTL, it ranged from 6.95 to 35.00 with a mean of 17.59 (Supplementary Table 1).

Among the GS methods, SVMRBF demonstrated lower predicted genotypic values at the response limits to and shorter half-lives (Figure 4). PS tends to result in larger values for response at the limits with longer half-lives. RREML, BayesB and BL methods demonstrated similar predicted genotypic values at the limits and half-lives, but showed significant variation depending on SI, TS and H factors (Figure 4). The modeled genotypic values were also highly correlated with the simulated genotypic values (Pearson correlation coefficient: 0.94-0.97).

4.2.2 Relative to PS, the three parametric GP models provided greater initial rates of response, reduced population half-life and faster loss of genetic variance. Selection using the SVMRBF consistently produced the least effective responses (Figure 5 and 6). Training sets consisting of data from up to 14 prior cycles of selection, compared to TS's consisting of data from only the

current cycle of selection are observably distinctive for both the short term and long term response metrics (Figure 5 and 6). Importantly, use of the TS's significantly improved responses to selection in both the short term and long term for all other combinations of factors (Supplementary Table 1). Selection intensity is the third most important factor to affect the response metrics (Supplementary Table 1). Consistent with theory (Robertson 1960; Hill and Robertson 1968; Bulmer 1971,1976), greater SI's (associated with retention of smaller proportions of RILs to initiate subsequent cycles, were associated with more rapid response rates, shorter half-lives, faster loss of genetic variance and significantly lower Rs values as the population approached it's selection response limits.

The maximum realized response for 40 simulated QTL was from 0.32 to 0.78 of maximum genotypic potential among the founders, whereas for 400 and 4289 QTL, the realized response was 0.12 - 0.42 and 0.04-0.17 of the maximum (Figure 5 and Supplementary Figure 15). Also if there are 40 simulated QTL, the maximum attained values are as high as 80% of the maximum value of 200 in less than ten cycles of recurrent selection (Supplementary Figure 16 & 17). In contrast, Rs values are no greater than 40% of the maximum value and stop responding to selection in 10-15 cycles if there are 400 simulated QTL. If nQTL = 4289, Rs values were never greater than 15% of the maximum value and do not begin to approach a response limit until after 20 cycles.

As expected, responses to selection are associated with declining genetic variances (Figure 7 and Supplementary Figure 18). The loss of Sgv's across cycles is much faster with fewer QTL than larger numbers of QTL. Likewise the estimated prediction accuracies (Figure 8 and Supplementary Figure 19) approach zero as the genotypic variance approaches zero. Average MSE for the GP models increase across cycles of selection (Figure 9 and

Supplementary Figure 20) and LD among markers approach zero as the genotypic variance approaches zero, although the covariance among these response metrics depend on the other simulated factors (Supplementary Figure 21-25).

To illustrate the impact of nQTL, consider Rs values plotted across forty cycles of recurrent selection (Figure 5 and Supplementary Figure 15). If the genetic architecture of the trait consists of 40 and 400 QTL, responses to selection were limited after 10-15 cycles of selection, whereas for 4289 QTL, limits to selection responses were not realized until 30 to 40 cycles of selection.

To interpret the role of nQTL as a factor, it is important to recall that: 1) positive and negative allelic effects were simulated to alternate sequentially at QTL (marker loci) that were distributed across the genome according to the Soybean genetic linkage map. 2) Crossing nearly homozygous lines and subsequently self-pollinating progeny before genotyping and phenotyping within each cycle creates a limited number of large linkage blocks. Analyses of the number of linkage blocks each generation reveals that regardless of the number of QTL, the number of linkage blocks per cycle ranges from 70-90. These were not the same blocks for each simulated set of lines each cycle, but if the nQTL equals 40, then each linkage block included all segregating QTL. For 400 and 4289 simulated QTL, each linkage block had a net genetic effect of zero or +/- 0.5 or +/- 0.05 multiplied by the number of QTL in the block respectively. Thus, the nQTL might be better understood as the magnitude of genetic effects associated with segregating linkage blocks.

The observable general trend for H, or perhaps more accurately understood as contributions of non-genetic effects to the phenotypes, was that H values of 0.7 for the initial phenotypic variance resulted in Rs values that were greater than H values of 0.3 (Figure 5 and

Supplementary Figure 15). The trend in Rs values is correlated with the other response metrics, in particular prediction accuracies of the GP models. The loss of estimated prediction accuracies are greater with H values of 0.3 than 0.7 with relaxed selection intensities (Figure 8 and Supplementary Figure 19). Other combinations of SI and H require model updating to provide reasonable prediction accuracies and achieve greater responses across more cycles of selection. For all combinations of SI and nQTL, losses of genotypic variance are greater with H values of 0.7 than 0.3 (Figure 7 and Supplementary Figure 18).

4.3 Some Specific Outcomes of Interest

For purposes of illustrating interaction effects on SM's across all 40 cycles of selection, consider the most relaxed SI of 1.75, associated with selecting 10% of the lines per cycle. When GP models are not updated, BB and BL produced greater R_s values than PS in the early cycles for all nQTL and both levels of H, whereas PS resulted in greater responses than all GS methods after the 10^{th} cycle (Figure 5). SVMRBF did not demonstrate any better responses than PS in either early or late cycles for any nQTL or level of H (Figure 5, Supplementary Figure 15, 26, and 27).

If the parametric GP models are updated with training sets consisting of data from up to 14 prior cycles of recurrent selection, responses to RR-REML demonstrated the greatest responses (Rs) for 40, 400 and 4289 QTL across both levels of H (Figure 5, Supplementary Figure 15 and 28). If the RR-REML model is updated with up to 14 prior cycles of training sets, responses are larger than PS for up to 10-40 cycles depending on the number of QTL, H and SI (Figure 5 and 6). When BB and BL GP models are updated, responses are larger than PS for up to 5, 20 and 40 cycles for 40, 400, and 4289 QTL respectively. Similar, albeit distinctive, comparisons among outcomes from GP models with model updating for genetic architectures

responsible for 0.3 of the phenotypic variance in the initial sets of RILs (Supplementary Figure 28) are described in Supplementary File 4.

Relative to responses without model updating application of RR-REML and Bayesian methods with model updating resulted in greater realization of the genetic potential of the founders. Model updating with Bayesian methods resulted in less favorable responses than the RR-REML. SVMRBF when updated with TS's demonstrated no significant improvement relative to SVMRBF without updating for all genetic architectures, levels of H and SI's (Supplementary Figure 27 and 28). If the genetic architecture explains only 30% of the phenotypic variability in the initial sets of lines, the relative improvements in Rs values across cycles using updated TS's are better than simulated QTL that explain 70% of the phenotypic variance (Supplementary Figure 28).

If GP models are not updated with data from up to 14 prior cycles, the Mgv's were consistently greater with PS than the four GS methods. Among GP models without updating, BB provided the best Mgv, while SVM-RBF had the smallest Mgv (Supplementary Figure 16 and 17). If GP models are updated, the pattern depends mostly on the number of QTL. For initial H values of either 0.7 and 0.3 and 40 simulated QTL, Mgv's are similar for RR-REML, Bayesian GP models and PS, whereas for 400 QTL, RR-REML produces greater Mgv's than PS and Bayesian methods. For 4289 QTL, RR-REML and Bayesian methods produce greater Mgv's than PS. Recurrent GS with SVMRBF produced the least desirable Mgv's for 40, 400 and 4289 QTL.

Outcomes for other combinations of factors: Percentage gains in responses for GS with model updating relative to response from PS for 400 QTL responsible for 30% of phenotypic variability is provided in Supplementary Figure 29. Percentage gains in responses for GS with model

updating relative to response from PS for all combinations of factors are provided in Supplementary File 5. Percentage gains in responses for GS with model updating relative to response from GS without model updating for 400 QTL responsible for 70% and 30% of phenotypic variability is provided in Supplementary Figure 30 and 31. Percentage gains in responses for GS with model updating relative to response from GS without model updating for all combinations of factors are provided in Supplementary File 6.

4.3.2 *Loss of Genotypic Potential and Variance* In terms of lost genetic potential, every cycle of selection reduced the maximum possible genotypic value. When GP models are not updated, the genetic potential is lost at a rapid rate beginning in the early cycles, whereas when GP models are updated, genetic potential is retained in the population and genetic variance decreases at a slower rate. PS had the least loss in genetic potential relative to all four GP models without updating. However, with model updating, the loss of genetic potential using parametric GP models was almost the same as PS. Among the parametric GP models, RR-REML and Bayesian methods showed similar slow losses of genetic potential with and without model updating. SVMRBF GS had the greatest loss of genetic potential beginning with the early cycles (Figure 6).

The loss of genetic potential in early cycles determines the limits to selection response in later cycles. For example, with 400 simulated QTL responsible for 70% of the phenotypic variance, the maximum potential was only 50% of the maximum potential (100 units) for PS and parametric GS methods with and without model updating. When parametric GP models are updated, 81% and 75% of the maximum available potential are realized with RR-REML and Bayesian methods respectively. If GP models are not updated, only 62% of the limits of the

maximum available potential are realized with RR-REML and Bayesian methods. With PS, 78% of maximum available potential is realized by the cycles in which the population no longer responds to selection. With SVMRBF, only 35 % of the potential is realized with and without updating by the cycles in which the population no longer responds to selection (Figure 6).

If GP models are updated, the standardized genotypic variance (Sgv) decreases at a rate similar to GP models that are not updated (Figure 7 and Supplementary Figure 18). There is no difference among GS methods in terms of rate at which Sgv decreases. Also, model updating significantly improved estimated prediction accuracies, $r_{ps,}$ for all GP models except SVMRBF. Among RR-REML and Bayesian GP models, model updating has a slightly larger impact on estimated accuracies and MSE using RR-REML than with Bayesian GP models (Figure 8, 9, Supplementary Figure 19 and 20). MSE were orders of magnitude lesser for RR-REML than bayesian GP models with updates after the first 10-15 cycles of selection (Figure 9 and Supplementary Figure 20).

If models are updated using data from up to 14 prior cycles, the changes to genetic variance among the RILs selected to be crossed, their average heterozygosity, average rate of inbreeding, and loss of favorable alleles are similar among GS methods (Supplementary Figure 32- 35). Model updating resulted in faster loss of genotypic variance among genotypes selected to be parental lines for the next cycle of inter-mating. The loss of genotypic variance is similar among parametric GS methods. (Supplementary Figure 32). When there are 400 simulated QTL responsible for 70% of the phenotypic variance, the average number of favorable alleles that are lost across 40 cycles due to selection and drift are similar among PS and GS methods, but the rate at which they are lost differs among selection methods for the first 20 cycles until they converge at the same limit. SVMRBF GS showed the greatest rate of loss and PS had the least

rate of loss while the parametric GS methods had intermediate rates of loss. By the time 40 cycles of selection have completed, model updating didn't result in any significant difference in rates of loss and the total number of favorable alleles that are lost (Supplementary Figure 33).

SVMRBF GS showed the greatest loss of average heterozygosity and PS lost heterozygosity at the slowest rate, while RR-REML and bayesian GS methods lost heterozygosity at intermediate rates. Model updating didn't result in significant changes to rates at which heterozygosity was lost over cycles (Supplementary Figure 34). PS showed slower rates of inbreeding than GS methods as we would expect from the decay of standardized genotypic variance. Average rates of inbreeding were similar among parametric GP models in the early cycles of recurrent selection, whereas the patterns varied after there is no response to selection (Supplementary Figure 35).

4.3.3 Response standardized to change in genotypic variance

The limiting values for RsVar (Response standardized to change in genotypic variance) when PS is used to select the best 10% of RILS with genetic architectures consisting of 400 and 4289 QTL are greater than the limiting values using GS methods without model updates (Supplementary Figure 36 and 37). The parametric GP models, without model updating, resulted in similar changes of RsVar for 40, 400 and 4289 simulated QTL responsible for both 70% and 30% of phenotypic variability in the initial population. Also, if the GP models are not updated, the rates and limits to loss of RsVar are similar among the GS methods for all nQTL and SI.

If GP models are updated with the TS's, the patterns of RsVar are significantly different among GS methods and are dependent on nQTL, SI and H (Supplementary Figure 36 and 37). With 0.7 heritability, there are no significant difference in RsVar among GS methods for 40 simulated QTL. If the genetic architecture consists of 400 and 4289 QTL and relaxed selection

intensities are practiced, the RR-REML GS method maintained genetic variance and RsVar for more cycles than PS and the Bayesian GS methods. Relative Gain in RsVar with RR-REML GS is even larger for 0.3 H treatment with relaxed selection intensities. SVMRBF demonstrated the least limits of RsVar for treatment combinations with and without model updating (Supplementary Figure 36 and 37). The plots for all the evaluation metrics discussed above for selection intensities 2.67 and 2.34 are available on request.

For all selection methods pairwise LD among markers on the same chromosome decreased across cycles of recurrent selection (Supplementary Figure 21-25). LD decreased slowest with PS (Supplementary Figure 21). Loss of LD in early and late cycles of selection are similar among parametric GP models and SVMRBF with the relaxed selection intensity. By the 20^{th} cycle of recurrent selection, LD approached zero for all selection methods and there was no evidence that selection methods affected linkage disequilibrium (LD) differentially in the earlier cycles. The rates at which LD decays are lower when GP models are updated with training sets compared to GP models without updating (Supplementary Figure 21-25).

4.4 Tradeoff for Short-term and Long-term Gains

For purposes of illustration, consider Rs values for 400 simulated QTL responsible for 70% of variability and the most relaxed SI of 1.75, associated with selecting 10% of the RILs per cycle. A weighted ranking method can be used to select the best GS method and training set combination for achieving short-term and long-term objectives. Considering that standardized responses or percent gain in response relative to a common reference is the objective to be maximized, it is possible to assign to relative weights to emphasize only the short-term or both short-term and long-term responses depending on the program objectives. Assuming a constant cycle time for GS with and without updates, RR-REML with model updating was the best method for both achieving both short-term and long-term objectives followed by BB and BL

with updates and PS (Table 3). However, when only the first ten cycles are emphasized, BL and RREML without model updates also demonstrated equivalent responses (Table 3). Provided that model updating requires additional cycles, GS method without updates could offer larger gains/cycle for some treatment combinations.

5. Discussion

5.1 General Discussion

Structures of plant populations are highly dependent on the reproductive biology. Indeed, plant breeders design genetic improvement projects based on reproductive biology. Similar to many cereal and pulse crops, the reproductive biology of soybean is primarily through self-pollination (Wilcox et al. 1979; Fehr 1980, 1991). The frequency of natural cross pollination in soybean is only about 0.025 (Garber et al. 1925; Caviness 1966; Carlson and Lersten 1987; Ahrent and Caviness 1994). Because crossing soybean lines is labor intensive and expensive soybean breeders use mating designs in which only one or two elite varieties are crossed with a few dozen recently selected lines (Guo et al. 2013; Guo et al. 2014). We refer to this as a hub network design. Soybean breeders subsequently take advantage of natural self-pollination to create lines with sufficient seed for replicated evaluations across many environments.

Our simulations attempted to emulate cycles of selection, crossing, and self-pollination currently conducted by commercial and academic soybean breeders. Relative to outcrossing species selecting homozygotes to participate in a hub network within each cycle will have smaller effective populations and retain LD for more cycles of recurrent selection. Future studies are needed to determine if the significant interaction effects that we found are due to population structure and genome organization.

Previous publications of *in silico* investigations of factors affecting outcomes from RGS have been conducted using a few factors applied to arbitrary diploid genomes, and an expected population structure from an assumed coalescent process. To our knowledge, the research reported herein is the first designed to reveal interactions among five factors that previously had been shown to affect responses to selection in populations created with existing contemporary soybean germplasm adapted to MZ's II and III. Our ability to detect and characterize interaction effects is enabled by use of a first order recurrence equation (eqn 4). Hopefully, our explanation of how to implement recurrence equation models in available R packages will encourage others to investigate recurrent genetic improvement designs. Our primary motivation for the use of non-linear modeling in this study was to provide systematic investigation of significance of variation in response to factors and detection of interaction effects. In the future we plan to use non-linear recurrence models to identify breeding strategies that will optimize outcomes with respect to competing objectives of maximizing responses to selection and retention of useful genetic variance.

5.2 Implications for Application of GS for Genetic Improvement in Soybeans

There has been considerable concern expressed about the limited genetic variability among soybean genotypes adapted to specific maturity zones and recurrently selected over 75 years in North America (Carter et al. 2004; Hyten et al. 2006; Mikel et al. 2010; McCouch et al. 2013). The SoyNAM founders represent a sample of contemporary breeding germplasm adapted to and used for agricultural production in maturity zones II - IV (Song et al. 2017; Diers et al. 2018). Despite concerns about limited genetic variability, our assessment of the half-lives suggest that even if only 40 QTL are segregating among founders of the SoyNAM panel, there is genetic potential for response to selection for at least five cycles. If there are larger numbers of

QTL with smaller additive effects distributed among linkage blocks of SoyNAM founders, then we can expect half-lives to RGS for at least 10 cycles (Supplementary File 2).

Among the five factors we investigated, a soybean breeder can choose SM's, SI's and TS's. Currently, plant breeders have little control on nQTL and H, because these parameters are determined by the sample of germplasm and environments. While it is possible to adjust the value of H on an entry mean basis by increasing/decreasing the number of replicates (Fehr 1991), estimating the number of segregating QTL and the magnitude of their effects is difficult and usually extremely expensive (Beavis 1994; Goring et al. 2001; Xu 2003). For a fixed budget, the breeder will be faced with a trade-off between numbers of replicates and numbers of lines that can be evaluated in the field. In other words, H on an entry mean basis can be estimated, but not adjusted without adding field plot resources. Also the nQTL and their contributions to genetic variance can be estimated, but the estimates will be biased.

Although modulated by nQTL and H, PS consistently retained useful genetic variability across many cycles of genetic improvement, Bayesian methods provided the fastest genetic gains in the short term and RR-REML provided a compromise between PS and Bayesian methods. SVM-RBF should not be considered for genetic improvement for the additive genetic architectures in soybean population structures (Table 3). Previously others have reported that long-term response using GS methods will be more limited than PS in closed populations (Goddard 2009; Zhong et al. 2009; Jannink 2010). We also observed that GS without updated TS's result in rapid loss of genetic variance in the initial cycles, which results in lower Rs values as responses to selection approach an asymptotic limit. When models were updated with TS's composed of data from prior cycles of selection, loss of prediction accuracy slowed for all values of nQTL and H (Figure 8).

Given similar rates of decreasing genetic variance among parametric GS methods, the different limits to selection response are likely due to greater efficiency of retaining genetic diversity with RR-REML for later cycles, although for many combinations of factors the limits of response with RR-REML and Bayesian methods are about the same. Unlike previous studies, we noted in the initial cycles of RGS genetic gains and estimates of accuracy were similar using BL and RR-REML, whereas after 15 cycles genetic responses were not as limited with RR-REML, probably because additive genetic variance was retained for later cycles of selection (Liu et al. 2015).

Replicated responses to high values of SI quickly reach a limit in five to ten cycles of recurrent selection. Replicated responses to lower values of selection intensity consistently result in greater gains over more cycles, indicating that genetic drift is the most likely mechanism for loss of genotypic variance. These constraints on plant breeding programs are well characterized (Brisbane and Gibson 1995; Hayes et al. 2009; Jannink 2010; Hung et al. 2012; Liu et al. 2015; Akdemir and Sánchez 2016; Yabe et al. 2016).

Model updating with TS's from prior cycles improves the relationship between training sets and validation sets and thus improves responses to GS without updated training sets. While model updating doesn't significantly change the estimated half-life, model updating did produce greater responses standardized to the rate at which genotypic variance is lost in selected populations with updated RR-REML GP models (Supplementary File 2; Supplementary Table 3).

The choice of which combinations of SM, SI and TS depend on the objectives of the breeding program. If the objective is to enter and capture market share in a short time, then maintenance of genetic diversity is not important. Averaged over nQTL and H, the greatest

changes in Rs values, i.e., rate of genetic gain, were attained in early cycles using BayesB and SI = 2.84, without model updating. Thus, if a soybean breeding project wants to maximize the rate of genetic gain for a single quantitative trait in a population derived from a sample of the SoyNAM founders, then application of the BayesB GS method and large SI is the best combination to meet the breeding objective. If the breeding project has a long term breeding objective to improve germplasm while maintaining useful genetic diversity for purposes such as providing useful germplasm for future generations or evaluating genome editing, then PS or GS with RR-REML with relaxed SI will be the best combination to meet the breeding objective. The greatest values for Rs were attained using the RR-REML model with model updating and SI = 1.75 (Table 3). A weighted ranking method for emphasizing short-term and long-term objectives could serve as a useful tool to determine the best SM and TS combination for each of the nQTL and H combinations. Last if the breeding project has multiple objectives for both immediate and longer term goals, then pareto optima among tradeoffs involving responses/cycle and retention of useful genetic variance for multiple traits need to be identified (Akdemir et al. 2019).

5.3 Lessons for Future Simulation Studies.

In our simulations we assigned the same amount of time to develop and evaluate lines for GS as for PS. Application of PS usually requires field trials for three years before lines are selected for use in a crossing nursery. In practice, one of the advantages of GS relative to PS is that only subsets of lines need to be phenotyped. Commercial soybean breeding projects have used genotypic values obtained from GP models to cull lines in as they are being self-pollinated and prior to the first stage of field trials (Kurek 2018). Also it is possible to train or update GP models with lines derived in earlier filial generations thereby requiring less time per cycle (Bassi et al. 2016). Even if both GS and PS require the same amount of time to develop lines before phenotypic evaluation, GS methods can be conducted as soon as phenotypic information is

available from first year field trials (Heffner et al. 2009), while PS usually requires phenotypic evaluations for three years before lines are selected for crossing nurseries to begin a new cycle. Herein we did not investigate the trade-offs between less accurate predictions from models trained with less extensive phenotyping or phenotyping with lines derived in earlier filial generations. These practical adjustments from application of GS methods to current soybean genetic improvement projects still need to be investigated.

Accuracies of GP models in the founding set of RILs were similar to that reported in previous studies (Long et al. 2010, 2011; Guo et al. 2012; Howard et al. 2014). However, our estimates of accuracies are larger as we've included QTL in our training sets. Relationship among selected RILs and LD between marker loci (ML) and QTL are considered the two most important sources of prediction accuracy. In previous studies, relationship among RILs had a greater effect on prediction accuracies for RR-REML than BL, whereas accuracy of BL was more dependent on LD and both components showed similar contribution to the accuracy of BayesB GP models (Habier et al. 2007; Zhong et al. 2009; Liu et al. 2015). Similar to our results without model updating, Bayesian GS methods resulted in greater responses as the populations approached asymptotic limits (Meuwissen 1997; Li et al. 2008; Akdemir and Sánchez 2016). However, GP model updating reduced the difference in rate of decrease of prediction accuracies among RR-REML and Bayesian GP models but there was no consistent pattern in relationship among selected RILs and rate of loss for LD to explain the observed estimates of prediction accuracies. However, GP model updating consistently resulted in lesser MSE for RR-REML than BB and BL GP models across all levels of SI, nQTL and H. This pattern is consistent with greater efficiency of converting loss of variance into gain with updated RR-REML GS method. In order to estimate the contribution of LD and linkage blocks to prediction accuracy of GP

models will require a design similar to that employed by Müller et al. (2017, 2018) for synthetic populations. Populations with unrelated training and prediction sets with LD and SNP based relationship estimates showed low prediction accuracy and low genetic response in recurrent GS similar to GS without updating in this study. Whereas populations with relationship between training and prediction sets with LD and SNP based relationship estimates showed greater prediction accuracy and greater genetic response similar to GS with model updating (Müller et al. 2017, 2018).

While TS's had relatively small impacts on response metrics, they were highly significant. Since relatedness of TS's and validation sets affect estimates of prediction accuracy it has been suggested that model updating needs to be evaluated for accuracies of prediction within and across populations (Crossa et al. 2014; Juliana et al. 2018; Stewart-Brown et al. 2019). Herein, the TS's were comprised of a combined population of RILs derived from all families across cycles of selection. Given a constant number of RILs in the TS's, within or across family prediction accuracies will depend on genetic differentiation among families (de Roos et al. 2009; Schulz-Streeck et al. 2012; Stewart-Brown et al. 2019). However, actual soybean breeding projects evaluate a few to many dozen RILs per family and future simulation studies, especially of two part systems, should consider whether relationships between evaluation units, RILs and selection units (Holland et al. 2003), possibly individuals, need to be used in design of the TS's.

We allowed the size of TS's to increase every cycle by adding data from prior cycles. Increasing the number of RILs per TS requires more computational resources. An alternative strategy is to randomly sample subsets of data from each of the prior cycles to maintain a constant cumulative training population size. It is also possible to assign weights to the samples

from prior cycles to place more weight on data from more recent cycles. These possibilities suggest determination of an optimal combination of numbers of RILs and weights that will provide maximum prediction accuracy with minimal computational requirements. Some aspects of this optimization problem have been addressed (Lorenz et al. 2013; Hickey et al. 2014; Akdemir et al. 2015; Xavier et al. 2017). For example, Akdemir et al (2015) devised a genetic algorithm for selecting optimal training populations to minimize prediction error variance and Xavier et al (2017) developed sampling methods for training Bayesian GP models. Another consideration is whether TS's need to be updated every cycle. Instead of updating the model with data from every cycle would it be more effective to retrain GP models every second third, fourth… cycle while maintaining a constant training population size? Modifications to design of TS's will need to be addressed with simulations before implementation.

Like most simulation studies we fixed levels of SI as constant for all combinations of factors across all cycles of recurrent selection. This is not consistent with actual soybean breeding projects. Just as sampling families result in some with exceptional genotypes and some with poor genotypes, in an actual genetic improvement project there is variability among cycles. The effects of applying a dynamic selection strategy is an alternative and interesting question. We hypothesize that a strategy consisting of applying different SI's, optimized for each cycle, will achieve improved long-term genetic response by differentially emphasizing genetic variance and genetic gain across multiple cycles of selection.

Unlike actual soybean genetic improvement projects we simulated a closed breeding population derived from a sample of SoyNAM founders in which culled lines were not resampled for discarded favorable alleles, nor did we simulate exchange of lines among breeding projects. In MZ's II, III and IV there are six public breeding projects and about a dozen

commercial soybean breeding projects. Thus, there is potential, depending on material transfer agreements, to exchange genotypes among breeding populations. In our next set of investigations migration among multiple breeding projects will be evaluated for response to recurrent selection within and among breeding projects using island model evolutionary algorithms (Hagan et al. 2012; Yabe et al. 2016). Results reported herein will provide comparators for assessing impacts of migration policies relative to the other factors that affect responses to RGS.

Also, as with prior simulation studies, we simulated truncation selection, but unlike previous investigations we did not randomly mate selected lines each cycle. Rather, we used the hub network design (Guo et al. 2013; Guo et al. 2014). We did not consider relationships among selected RILs nor the trade-offs between genetic gain, genetic variance (inbreeding) when selecting RILs to cross. There exist multi-objective optimization methods such as genomic mating and optimal cross selection (Rutten et al. 2002; Woolliams et al. 2015; Akdemir and Sánchez 2016; Gorjanc et al. 2018; Akdemir et al. 2019) that have been demonstrated to provide both greater rates of genetic gain and assure maintenance of population genetic potential across cycles.

In the future, it is possible that development of male sterile and insect pollination systems for soybean (Ortiz-Perez 2008; Davis 2020) will enable a cycle of RGS to be conducted using two or three mating generations per year. This will enable an alternative genetic improvement system based on decoupling genetic improvement from variety development (Gaynor et al. 2017). By separating the two types of breeding projects, GS can be applied continuously. In the two part system TS's will need to be composed of genotypic and phenotypic data obtained from

annual field trials of RILs, although the TS's will be several selection cycles removed from the cycle used to create the RILs.

Results reported herein suggest that RGS in a two part system will rapidly produce genetic gains and loss of useful genetic variance in very short periods of time. Indeed, with emerging technologies that will make it easier to intercross and obtain doubled haploid lines, 40 cycles of RGS could be completed in 12 to 20 years. To offset the shorter cycle times, application of GP models to select and cross individual progeny instead of RILs could result in a larger effective population size per cycle by creating more opportunities for recombination and slow the unintended loss of valuable alleles in discarded linkage blocks. However, before investments in development of male sterile and insect pollination systems for soybean (Ortiz-Perez 2008; Byrum and Davis, 2014) research using simulations are needed to understand the trade-offs and whether the investments are justified using approaches from Operations Research (Xu et al. 2011; Cameron et al. 2017, Han et al. 2017).

Acknowledgments

Funding for this research was provided by the Department of Agronomy, Iowa State University, the North Central Soybean Research Program and an NSF grant (1830478). Supplemental funding for large scale computing was enabled by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation. XSEDE resources consisted of research allocations (DMS190015 & DMS190018) on PSC-Bridges Large Memory nodes for the simulations involving updating of parametric GP models. The Iowa State University-Pronto GPU cluster enabled computation of SVM model update simulations. We want to acknowledge Matheus de Krause for discussions on fitting non-linear models using 'nlme' package, Dr. Lizhi Wang for efficient programs to simulate meiosis

and Dr. Alencar Xavier for sharing an efficient expectation maximization method for fitting ridge regression GP models. Last, we want to thank anonymous critical reviewers of earlier versions of this manuscript for providing useful suggestions on how to present such a large volume of information from such a comprehensive investigation of expected short and long term outcomes of GS applied to soybeans adapted to MZ's II to IV. A preprint of a version of this article is available on biorXiv at https://www.biorxiv.org/content/10.1101/2020.02.14.949008v2

References

Achard F, Butruille M, Madjarac S, Nelson PT, Duesing J, Laffont JL, Nelson B, Xiong J, Mikel MA, Smith JSC: **Single nucleotide polymorphisms facilitate dus testing of soybean cultivars for plant variety protection.** *Crop Science* 2020, **n/a**.

Ahrent DK, Caviness CE: **Natural Cross-Pollination of Twelve Soybean Cultivars in Arkansas.** *Crop Science* 1994, **34:**376-378.

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J: **Multi-objective optimized genomic breeding strategies for sustainable food improvement.** *Heredity* 2019, **122:**672-683.

Akdemir D, Sánchez JI: **Efficient Breeding by Genomic Mating.** *Frontiers in Genetics* 2016, **7**.

Akdemir D, Sanchez JI, Jannink JL: **Optimization of genomic selection training populations with a genetic algorithm.** *Genet Sel Evol* 2015, **47:**38.

Anderson EJ, Ali ML, Beavis WD, Chen P, Clemente TE, Diers BW, Graef GL, Grassini P, Hyten DL, McHale LK, et al: **Soybean [Glycine max (L.) Merr.] Breeding: History, Improvement, Production and Future Opportunities.** In *Advances in Plant Breeding Strategies: Legumes: Volume 7.* Edited by Al-Khayri JM, Jain SM, Johnson DV. Cham: Springer International Publishing; 2019: 431-516

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L: **Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats.** Iowa State University Digital Repository; 2011.

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J: **Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.).** *Plant Science* 2016, **242:**23-36.

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H: **Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures.** *Genetics Selection Evolution* 2012, **44:**3.

Baty F, Ritz C, Charles S, Brutsche M, Flandrois J-P, Delignette-Muller M-L: **A Toolbox for Nonlinear Regression in R: The Package nlstools.** *Journal of Statistical Software; Vol 1, Issue 5 (2015)* 2015.

Beavis WD: **The power and deceit of QTL experiments: lessons from comparative QTL studies.** In *Proceedings of the forty-ninth annual corn and sorghum industry research conference*. Chicago, IL; 1994: 250-266.

Bernardo R: **Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years.** *Crop Science* 2008, **48:**1649.

Bernardo R: **Genomewide Selection of Parental Inbreds: Classes of Loci and Virtual Biparental Populations.** *Crop Science* 2014, **54:**2586.

Bernardo R, Yu J: **Prospects for Genomewide Selection for Quantitative Traits in Maize.** *Crop Science* 2007, **47:**1082-1090.

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, et al: **Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress.(RESEARCH)(Author abstract).** 2015, **55:**154.

Bijma P: **Long-term genomic improvement – new challenges for population genetics.** *Journal of Animal Breeding and Genetics* 2012, **129:**1-2.

Brisbane J, Gibson J: **Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions.** *International Journal of Plant Breeding Research* 1995, **91:**421-431.

Brooks C, Nekrasov V, Lippman ZB, Van Eck J: **Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system.** *Plant Physiol* 2014, **166:**1292-1297.

Bulmer MG: **The Effect of Selection on Genetic Variability.** *The American Naturalist* 1971, **105:**201-211.

Bulmer MG: **The effect of selection on genetic variability: a simulation study.** *Genet Res* 1976, **28:**101-117.

Byrum J, Beavis B, Davis C, Doonan G, Doubler T, Kaster V, Mowers R, Parry S: **Genetic Gain Performance Metric Accelerates Agricultural Productivity.** *Interfaces* 2017, **47:**442-453.

Cameron JN, Han Y, Wang L, Beavis WD: **Systematic design for trait introgression projects.** *Theoretical and Applied Genetics* 2017, **130:**1993-2004.

Cannon SB, Shoemaker RC: **Evolutionary and comparative analyses of the soybean genome.** *Breeding science* 2012, **61:**437-444.

Carlson JB, Lersten NR: **Reproductive Morphology. In Soybeans: Improvement, Production, and Uses.** 2004**:**59-95.

Carter Jr TE, Nelson RL, Sneller CH, Cui Z: **Genetic diversity in soybean.** *Soybeans: Improvement, production, and uses* 2004, **16:**303-416.

Caviness CE: **Estimates of Natural Crosspollination in Jackson Soybeans in Arkansas1.** *Crop Science* 1966, **6:**cropsci1966.0011183X000600020034x.

Cockerham CC, Burrows PM: **Selection limits and strategies.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77:**546.

Collins LM, Dziak JJ, Kugler KC, Trail JB: **Factorial Experiments: Efficient Tools for Evaluation of Intervention Components: Efficient Tools for Evaluation of Intervention Components.** *American Journal of Preventive Medicine* 2014, **47:**498-504.

Cooper M, Podlich D, Micallef K, Smith O, Jensen N, Chapman S, Kruger N: **Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops.** *Quantitative genetics, genomics and plant breeding'(Ed MS Kang) pp* 2002**:**143-166.

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, et al: **Genomic prediction in CIMMYT maize and wheat breeding programs.** *Heredity* 2014, **112:**48-60.

Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, Hickey JM: **Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking.** *Genetics* 2013, **193:**347-365.

De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J: **Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods.** *Genetics Research* 2010, **92:**295-308.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL: **Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding.** *Genetics* 2013, **193:**327-345.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: **Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree.** *Genetics* 2009, **182:**375-385.

de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D: **Prediction of complex human traits using the genomic best linear unbiased predictor.** *PLoS Genet* 2013, **9:**e1003608.

de Roos APW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations.** *Genetics* 2009, **183:**1545-1553.

Dempfle L: **A note on increasing the limit of selection through selection within families.** *Genet Res* 1974, **24:**127-135.

Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, et al: **Genetic Architecture of Soybean Yield and Agronomic Traits.** *G3: Genes|Genomes|Genetics* 2018, **8:**3367.

Dunn K: **Process Improvement using Data.** In.; 2015

Emily C, Rex B: **Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers.** *The Plant Genome* 2013, **6**.

Endelman JB: **Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP.** *The Plant Genome Journal* 2011, **4:**250.

Fehr W: *Principles of cultivar development: theory and technique.* Macmillian Publishing Company; 1991.

Fehr WR, Hadley HH: *Hybridization of crop plants.* American Society of Agronomy and Crop Science Society of America; 1980.

Forsberg SK, Bloom JS, Sadhu MJ, Kruglyak L, Carlborg Ö: **Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast.** *Nat Genet* 2017, **49:**497-503.

Garber RJ, Odland T, McIlvaine T, Quisenberry K: *Varietal Experiments with Soybeans.* Agricultural Experiment Station, College of Agriculture, West Virginia …; 1925.

Gaynor RC, Gorjanc G, Bentley AR, Ober ES, Howell P, Jackson R, Mackay IJ, Hickey JM: **A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines.** *Crop Science* 2017, **57:**2372-2386.

Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245-257.

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS: **Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection.** *Genetics* 2017, **206:**1675.

Goldberg S: *Introduction to difference equations, with illustrative examples from economics, psychology, and sociology.* New York: New York, Wiley; 1958.

Göring HH, Terwilliger JD, Blangero J: **Large upward bias in estimation of locus-specific effects from genomewide scans.** *The American Journal of Human Genetics* 2001, **69:**1357-1369.

Gorjanc G, Gaynor RC, Hickey JM: **Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection.** *Theoretical and Applied Genetics* 2018, **131:**1953-1966.

GOUDET J: **hierfstat, a package for r to compute and test hierarchical F-statistics.** *Molecular Ecology Notes* 2005, **5:**184-186.

Grant D, Cregan P, Shoemaker RC: **Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis.** *Proceedings of the National Academy of Sciences* 2000, **97:**4168-4173.

Grant D, Nelson RT, Cannon SB, Shoemaker RC: **SoyBase, the USDA-ARS soybean genetics and genomics database.** *Nucleic acids research* 2010, **38:**D843.

Guo B, Sleper DA, Beavis WD: **Nested Association Mapping for Identification of Functional Markers.** *Genetics* 2010, **186:**373-383.

Guo B, Wang D, Guo Z, Beavis WD: **Family-based association mapping in crop species.** *Theoretical and Applied Genetics* 2013, **126:**1419-1430.

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G: **The impact of population structure on genomic prediction in stratified populations.** *Theoretical and Applied Genetics* 2014, **127:**749-762.

Guo Z, Tucker DM, Lu J, Kishore V, Gay G: **Evaluation of genome-wide selection efficiency in maize nested association mapping populations.** *Theoretical and Applied Genetics* 2012, **124:**261-275.

Guzman C, Peña RJ, Singh R, Autrique E, Dreisigacker S, Crossa J, Rutkoski J, Poland J, Battenfield S: **Wheat quality improvement at CIMMYT and the use of genomic selection on it.** *Applied & Translational Genomics* 2016, **11:**3-8.

Habier D, Fernando RL, Dekkers JCM: **The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values.** *Genetics* 2007, **177:**2389.

Hagan S, Knowles J, Kell DB: **Exploiting Genomic Knowledge in Optimising Molecular Breeding Programmes: Algorithms from Evolutionary Computing (Evolutionary Computing for Molecular Breeding).** 2012, **7:**e48862.

Han Y, Cameron JN, Wang L, Beavis WD: **The Predicted Cross Value for Genetic Introgression of Multiple Alleles.** *Genetics* 2017, **205:**1409.

Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix.** *Genetics Research* 2009, **91:**47-60.

Heffner EL, Sorrells ME, Jannink J-L: **Genomic Selection for Crop Improvement.** *Crop Science* 2009, **49:**1.

Heslot N, Akdemir D, Sorrells ME, Jannink J-L: **Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions.** *Theoretical and Applied Genetics* 2014, **127:**463-480.

Heslot N, Jannink J-L, Sorrells ME: **Perspectives for Genomic Selection Applications and Research in Plants.** *Crop Science* 2015, **55:**1-12.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Science* 2012, **52:**146-160.

Hickey JM, Chiurugwi T, Mackay I, Powell W: **Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery.** *Nature genetics* 2017, **49:**1297.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G: **Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation.** *Crop Science* 2014, **54:**1476-1488.

Hickey JM, Gorjanc G: **Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods.** *G3 (Bethesda, Md)* 2012, **2:**425.

Hill WG, Robertson A: **The effect of linkage on limits to artificial selection.** *Genetics Research* 2008, **89:**311-336. (First published in 1968)

Holland J, Nyquist WE, Cervantes-Martinez CT: **Estimating and interpreting heritability for plant breeding: Anupdate.** *Plant breeding reviews* 2003, **22:**9-111.

Hou J, van Leeuwen J, Andrews BJ, Boone C: **Genetic Network Complexity Shapes Background-Dependent Phenotypic Expression.** *Trends in genetics : TIG* 2018, **34:**578-586.

Howard R, Carriquiry AL, Beavis WD: **Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures.** *G3 (Bethesda)* 2014, **4:**1027-1046.

Howard R, Carriquiry AL, Beavis WD: **Application of Response Surface Methods To Determine Conditions for Optimal Genomic Prediction.** *G3: Genes|Genomes|Genetics* 2017, **7:**3103.

Hung HY, Browne C, Guill K, Coles N, Eller M, Garcia A, Lepak N, Melia-Hancock S, Oropeza-Rosas M, Salvo S, et al: **The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population.** *Heredity (Edinb)* 2012, **108:**490-499.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB: **Impacts of genetic bottlenecks on soybean genome diversity.** *Proc Natl Acad Sci U S A* 2006, **103:**16666-16671.

Jannink J-L: **Dynamics of long-term genomic selection.** *Genetics Selection Evolution* 2010, **42:**35.

Jannink JL, Lorenz AJ, Iwata H: **Genomic selection in plant breeding: from theory to practice.** *Brief Funct Genomics* 2010, **9:**166-177.

Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, Hickey JM: **Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs.** *Genetics Selection Evolution* 2015, **47:**55.

Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers.** *Bioinformatics* 2008, **24:**1403-1405.

Jombart T, Ahmed I: **adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.** *Bioinformatics* 2011, **27:**3070-3071.

Jonas E, de Koning D-J: **Does genomic selection have a future in plant breeding?** *Trends in Biotechnology* 2013, **31:**497-504.

Jonas E, de Koning DJ: **Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops.** *Biotechnology & genetic engineering reviews* 2016, **32:**18.

Juliana P, Singh RP, Poland J, Mondal S, Crossa J, Montesinos-López OA, Dreisigacker S, Pérez-Rodríguez P, Huerta-Espino J, Crespo-Herrera L, Govindan V: **Prospects and Challenges of Applied Genomic Selection—A New Paradigm in Breeding for Grain Yield in Bread Wheat.** *The Plant Genome* 2018, **11:**180017.

Kang H: **Limits of artificial selection under balanced mating systems.** 1979.

Kang H: **Limits of artificial selection under balanced mating systems with family selection.** *Silvae genetica* 1983, **32:**188-195.

Kang H, Namkoong G: **Limits of artificial selection under unbalanced mating systems.** *Theoretical and Applied Genetics* 1980, **58:**181-191.

Kang H, Nienstaedt H: **Managing long-term tree breeding stock.** *Silvae genetica* 1987, **36:**30-39.

Kurek A: **"Phenotypic and genomic selection for multi-trait improvement in soybean line and variety development"** 2018.

Langewisch T, Lenis J, Jiang G-L, Wang D, Pantalone V, Bilyeu K: **The development and use of a molecular model for soybean maturity groups.** *BMC Plant Biology* 2017, **17:**91.

Lemmon ZH, Reem NT, Dalrymple J, Soyk S, Swartwood KE, Rodriguez-Leal D, Van Eck J, Lippman ZB: **Rapid improvement of domestication traits in an orphan crop by genome editing.** *Nat Plants* 2018, **4:**766-770.

Li Y, Kadarmideen HN, Dekkers JCM: **Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit.** *Journal of Animal Breeding and Genetics* 2008, **125:**320-329.

Liu H, Meuwissen TH, Sorensen AC, Berg P: **Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs.** *Genet Sel Evol* 2015, **47:**19.

Long N, Gianola D, Rosa GJM, Weigel KA: **Application of support vector regression to genome-assisted prediction of quantitative traits.** *Theoretical and Applied Genetics* 2011, **123:**1065.

Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, GonzÁLez-Recio O: **Radial basis function regression methods for predicting quantitative traits using SNP markers.** *Genetics Research* 2010, **92:**209-225.

Lorenz AJ: **Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment.** *G3: Genes|Genomes|Genetics* 2013, **3:**481.

Marulanda J, Mi X, Melchinger A, Xu J-L, Würschum T, Longin C: **Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale.** *Theor Appl Genet* 2016, **129:**1901-1913.

McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G: **Feeding the future.** *Nature* 2013, **499:**23-24.

Meuwissen T, Hayes B, Goddard M: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819-1829.

Meuwissen T, Hayes B, Goddard M: **Accelerating improvement of livestock with genomic selection.** *Annu Rev Anim Biosci* 2013, **1:**221-237.

Meuwissen TH: **Maximizing the response of selection with a predefined rate of inbreeding.** *Journal of animal science* 1997, **75:**934-940.

Meuwissen TH, Goddard ME: **Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data.** *Genetics Selection Evolution* 2004, **36:**261.

Michel S, Ametz C, Gungor H, Epure D, Grausgruber H, Loschenberger F, Buerstmayr H: **Genomic selection across multiple breeding cycles in applied bread wheat breeding.(Original Article)(Report).** *Theoretical and Applied Genetics* 2016, **129:**1179.

Mikel MA, Diers BW, Nelson RL, Smith HH: **Genetic Diversity and Agronomic Improvement of North American Soybean Germplasm.** *Crop Science* 2010, **50:**1219-1229.

Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, Juliana P, Singh R: **A Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding.** *G3: Genes|Genomes|Genetics* 2019, **9:**601.

Müller D, Schopp P, Melchinger AE: **Persistency of Prediction Accuracy and Genetic Gain in Synthetic Populations Under Recurrent Genomic Selection.** *G3: Genes|Genomes|Genetics* 2017, **7:**801.

Müller D, Schopp P, Melchinger AE: **Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection.** *G3: Genes|Genomes|Genetics* 2018, **8:**1173.

Myers RH: *Response surface methodology / Raymond H. Myers.* s.l. : [Ann Arbor: s.l. : s.n. Ann Arbor : available from Edwards Brothers; 1976.

Nakaya A, Isobe SN: **Will genomic selection be a practical method for plant breeding?** *Annals of Botany* 2012, **110:**1303-1316.

Norman A, Taylor J, Edwards J, Kuchel H: **Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy.** *G3: Genes|Genomes|Genetics* 2018.

Oddi FJ, Miguez FE, Ghermandi L, Bianchi LO, Garibaldi LA: **A nonlinear mixed-effects modeling approach for ecological data: Using temporal dynamics of vegetation moisture as an example.** *Ecology and evolution* 2019, **9:**10225-10240.

Orf JH: **2 - Breeding, Genetics, and Production of Soybeans.** In *Soybeans.* Edited by Johnson LA, White PJ, Galloway R: AOCS Press; 2008: 33-65

Ortiz-Perez E, Mian RMA, Cooper RL, Mendiola T, Tew J, Horner HT, Hanlin SJ, Palmer RG: **Seed-set evaluation of four male-sterile, female-fertile soybean lines using alfalfa leafcutting bees and honey bees as pollinators.** *The Journal of Agricultural Science* 2008, **146:**461-469.

Peccoud J, Velden KV, Podlich D, Winkler C, Arthur L, Cooper M: **The selective values of alleles in a molecular network model are context dependent.** *Genetics* 2004, **166:**1715-1725.

Pérez P, de los Campos G: **Genome-Wide Regression and Prediction with the BGLR Statistical Package.** *Genetics* 2014, **198:**483-495.

Pinheiro JC: *Mixed-effects models in S and S-PLUS / José C. Pinheiro, Douglas M. Bates.* New York: New York : Springer; 2000.

Podlich DW, Cooper M: **Modelling Plant Breeding Programs as Search Strategies on a Complex Response Surface.** In *Simulated Evolution and Learning: Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL'98 Canberra, Australia, November 24–27, 1998 Selected Papers.* Edited by McKay B, Yao X, Newton CS, Kim J-H, Furuhashi T. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999: 171-178

Robertson A: **A Theory of Limits in Artificial Selection.** *Proceedings of the Royal Society of London Series B Biological Sciences* 1960, **153:**234.

Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB: **Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing.** *Cell* 2017, **171:**470-480.e478.

Rutten MJM, Bijma P, Woolliams JA, van Arendonk JAM: **SelAction: Software to Predict Selection Response and Rate of Inbreeding in Livestock Breeding Programs.** *Journal of Heredity* 2002, **93:**456-458.

Ryman N, Leimar O: **GST is still a useful measure of genetic differentiation — a comment on Jost's D.** *Molecular Ecology* 2009, **18:**2084-2087.

Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP: **Genomic Selection using Multiple Populations.** *Crop Science* 2012, **52:**2453-2461.

Shoemaker R, Polzin K, Labate J, Specht J, Brummer E, Olson T, Young N, Concibido V, Wilcox J, Tamulonis J: **Genome duplication in soybean (Glycine subgenus soja).** *Genetics* 1996, **144:**329-338.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB: **Fingerprinting Soybean Germplasm and Its Utility in Genomic Research.** *G3: Genes|Genomes|Genetics* 2015, **5:**1999.

Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, Song B-H, Charles An Y-Q, Hyten D, Nelson R, et al: **Genetic Characterization of the Soybean Nested Association Mapping Population.** *The Plant Genome* 2017, **10**.

Specht JE, Diers BW, Nelson RL, de Toledo JFF, Torrion JA, Grassini P: **Soybean.** *Yield gains in major US field crops* 2014, **33:**311-355.

Stewart-Brown BB, Song Q, Vaughn JN, Li Z: **Genomic Selection for Yield and Seed Composition Traits Within an Applied Soybean Breeding Program.** *G3: Genes|Genomes|Genetics* 2019, **9:**2253.

Wang Z, Chu T, A Choate L, Danko C: *Rgtsvm: Support Vector Machines on a GPU in R.* 2017.

Weir BS: **Disequilibrium.** *methods for discrete population genetic data* 1996**:**91-139.

Wilcox JR, Schapaugh Jr. WT, Bernard RL, Cooper RL, Fehr WR, Niehaus MH: **Genetic Improvement of Soybeans in the Midwest1.** *Crop Science* 1979, **19:**cropsci1979.0011183X001900060014x.

Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C: **Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection.** *Genetics* 2013, **195:**573.

Woolliams J, Corbin L: **Coalescence theory in livestock breeding.** *Journal of Animal Breeding and Genetics* 2012, **129:**255-256.

Woolliams JA, Berg P, Dagnachew BS, Meuwissen TH: **Genetic contributions and their optimization.** *J Anim Breed Genet* 2015, **132:**89-99.

Xavier A: **Efficient Estimation of Marker Effects in Plant Breeding.** *G3: Genes|Genomes|Genetics* 2019, **9:**3855.

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan P, et al: **Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population.** *G3: Genes|Genomes|Genetics* 2017.

Xavier A, Xu S, Muir W, Rainey KM: **Genomic prediction using subsampling.** *BMC Bioinformatics* 2017, **18:**191.

Xavier A, Thapa R, Muir WM, Rainey KM: **Population and quantitative genomic properties of the USDA soybean germplasm collection.** *Plant Genetic Resources* 2018**:**1-11.

Xavier A, Muir WM, Rainey KM: **Assessing Predictive Properties of Genome-Wide Selection in Soybeans.** *G3: Genes|Genomes|Genetics* 2016, **6:**2611.

Xu S: **Theoretical basis of the Beavis effect.** *Genetics* 2003, **165:**2259-2268.

Xu P, Wang L, Beavis WD: **An optimization approach to gene stacking.** *European Journal of Operational Research* 2011, **214:**168-178.

Yabe S, Yamasaki M, Ebana K, Hayashi T, Iwata H: **Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops.** *PLoS One* 2016, **11:**e0153945.

Yu J, Holland JB, McMullen MD, Buckler ES: **Genetic Design and Statistical Power of Nested Association Mapping in Maize.** *Genetics* 2008, **178:**539.

Zhong S, Dekkers JCM, Fernando RL, Jannink J-L: **Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study.** *Genetics* 2009, **182:**355-364.

Zuur A: ***Mixed Effects Models and Extensions in Ecology with R*** *by Alain Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith.* 1st ed. 2009.. edn: New York, NY : Springer New York : Imprint: Springer; 2009.

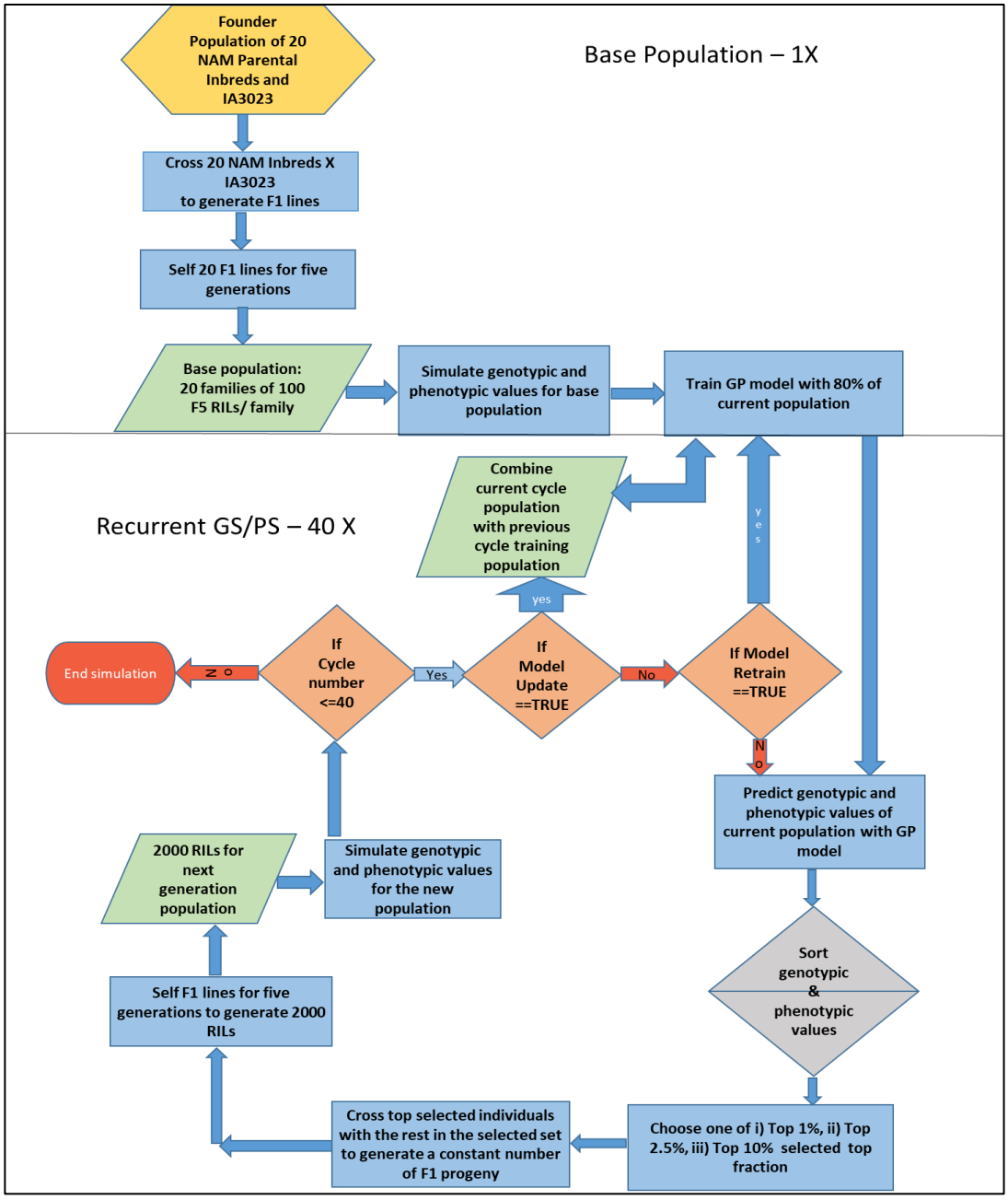

Figure 1 Flow Chart for Simulations of Recurrent Genomic Selection. The upper half panel represents the steps involved in generating the base population of 2000 F_5 RILs derived from 20 NAM founder lines crossed, *in silico*, to IA3023. It includes the model training step for genomic prediction models. The lower half panel represents recurrent steps of prediction, sorting, truncation selection, crossing, and generation of F_5 RILs for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process is to be continued for another cycle.

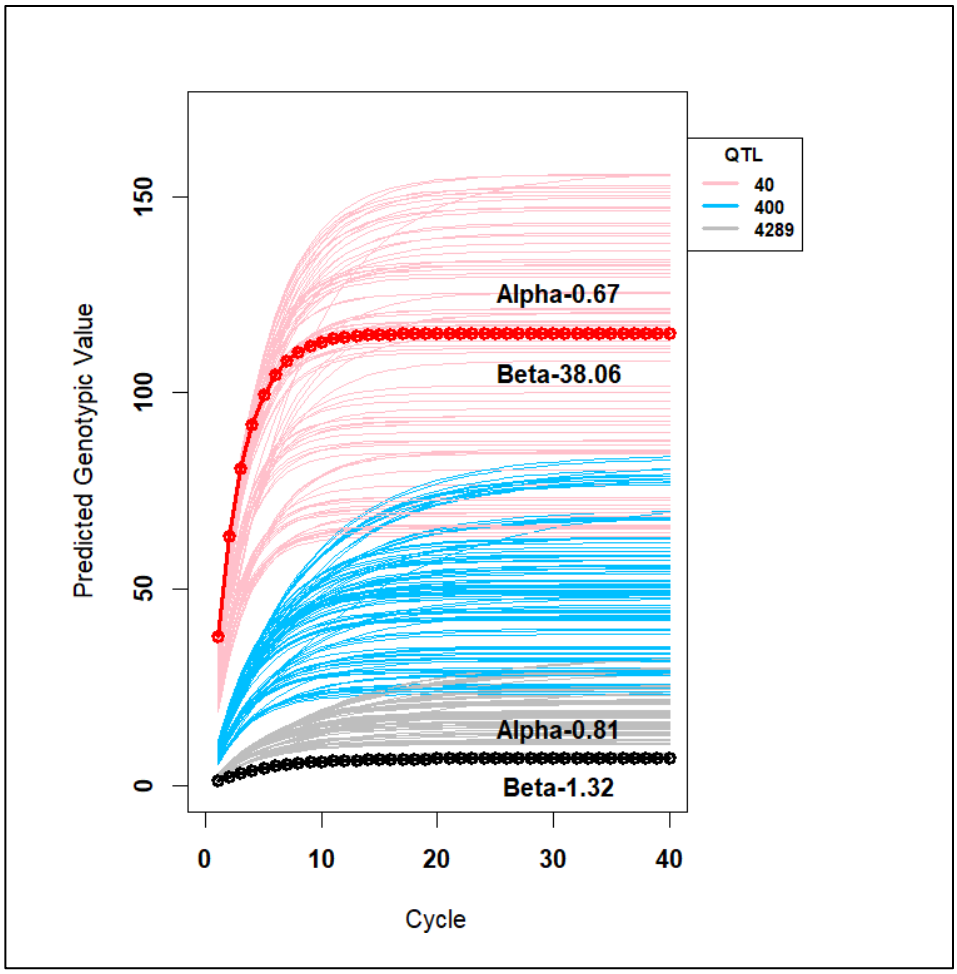

Figure 2. Predicted Genotypic Values from 40 cycles of recurrent selection modeled with the recurrence equation, where y_c represents the genotypic value in cycle c, with c= 0, 1..39 and values of α and β range from 0.6-0.9 and 1.32-38.06 respectively for 360 combinations of factors across all selection methods, training sets, selection intensities, number of simulated QTL and simulated heritabilities. The simulated QTL were distributed evenly throughout the genome, and each contributed equal additive effects of 5/-5, 0.5/-0.5, or 0.05/-0.05 units when there are 40, 400 and 4289 QTL respectively to the total genotypic value. The bold lines represent curves with the smallest and largest beta values and their corresponding alpha values. The curves are colored corresponding to 40 (pink), 400 (blue) and 4289 (gray) simulated QTL.

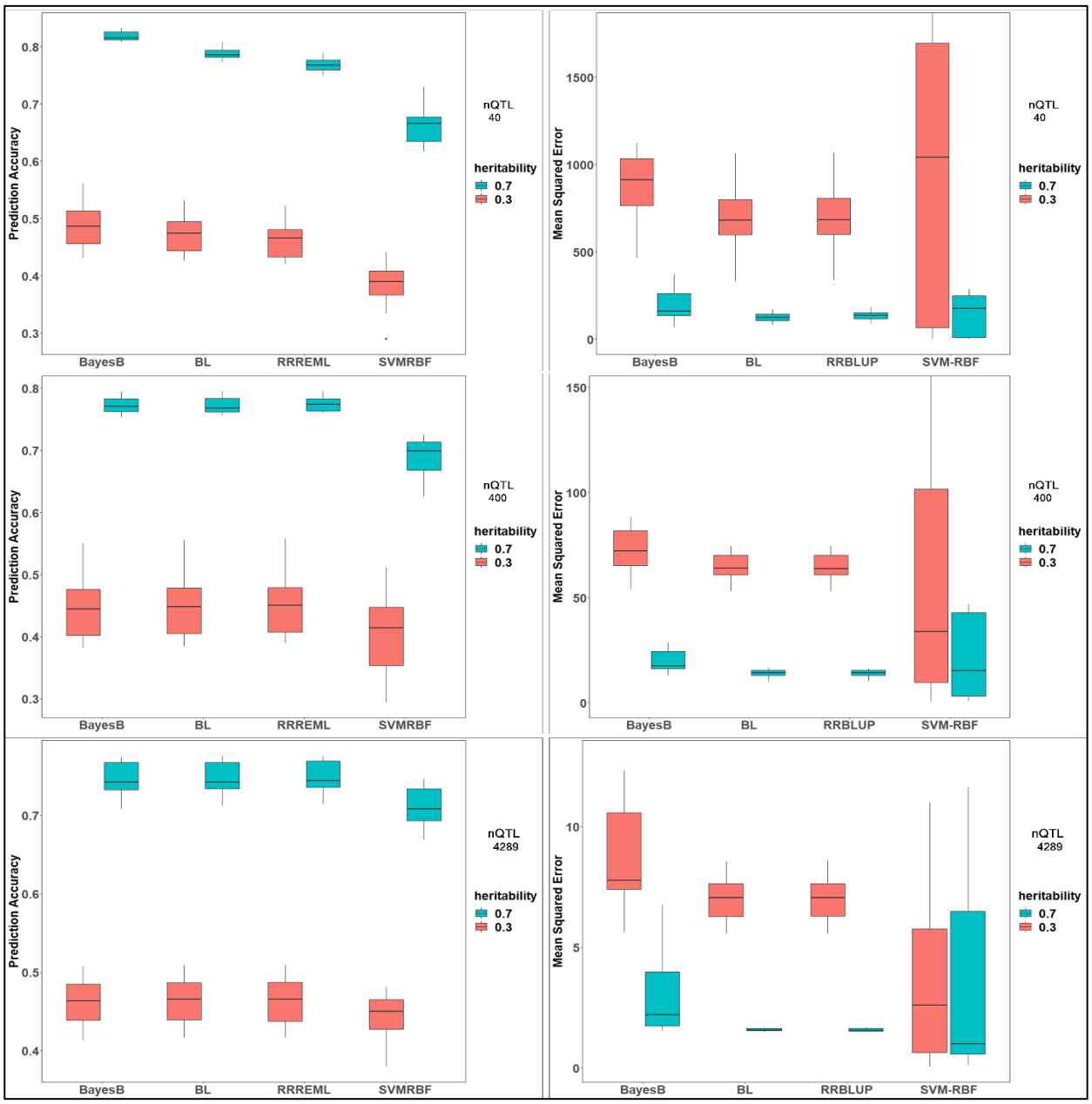

Figure 3 Estimated prediction accuracies and MSE in Founding Set of RILs: Estimated prediction accuracies (left panel) and mean squared errors (right panel) for four genomic prediction (GP) models: BayesB, BL (Bayes LASSO), RRREML (Ridge Regression with REML) and SVMRBF (Support Vector Machines with Radial Basis Function Kernel) trained with F_5 RILs derived from crosses of 20 homozygous founder lines with IA3023. Phenotypes used to train the GP models consisted of genetic architectures comprised of 40, 400 and 4289 simulated QTL (top, middle and bottom) that were responsible for 70% (blue) and 30% (red) of phenotypic variability in the initial populations.

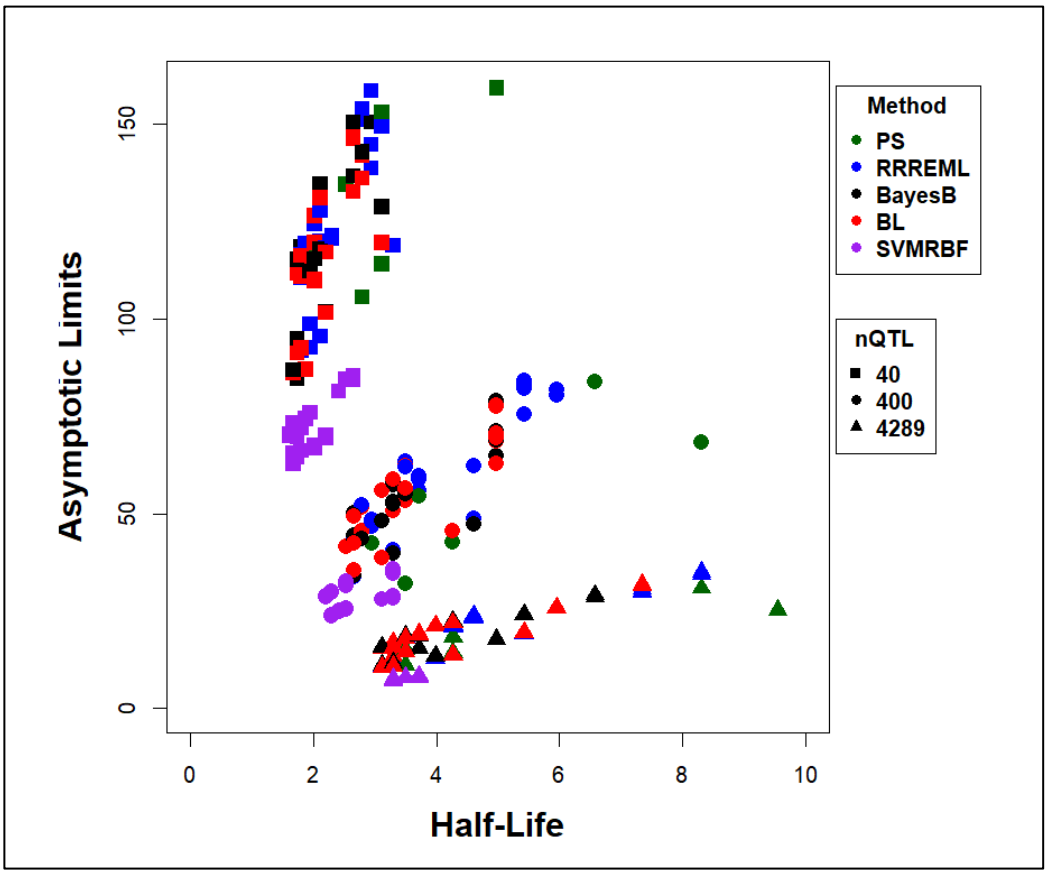

Figure 4 Asymptotic limits and Half-life of Recurrent Selection for the 360 combinations of factors. Half-life is plotted on the x-axis and asymptotic limits on the y-axis. The three major cluster of points correspond to 40 simulated QTL (square points), 400 simulated QTL (circular points) and 4289 simulated QTL (triangular points). Half-life and asymptotic limits vary depending on the combination of SM, TS, SI, nQTL and H factors. The selection methods include PS-Phenotypic Selection (green), RR-REML - Ridge Regression with Restricted Maximum Likelihood (blue), BayesB (black), BL – Bayes LASSO (red), and SVMRBF- Support Vector Machine with Radial Basis Kernel (purple). Selection intensities include top 1%, 2.5% and 10% selected fraction. Training sets include '0', which corresponds to no updating, and model updating with up to 10, 12 and 14 prior cycles of training data. H includes 0.7 and 0.3 heritabilities

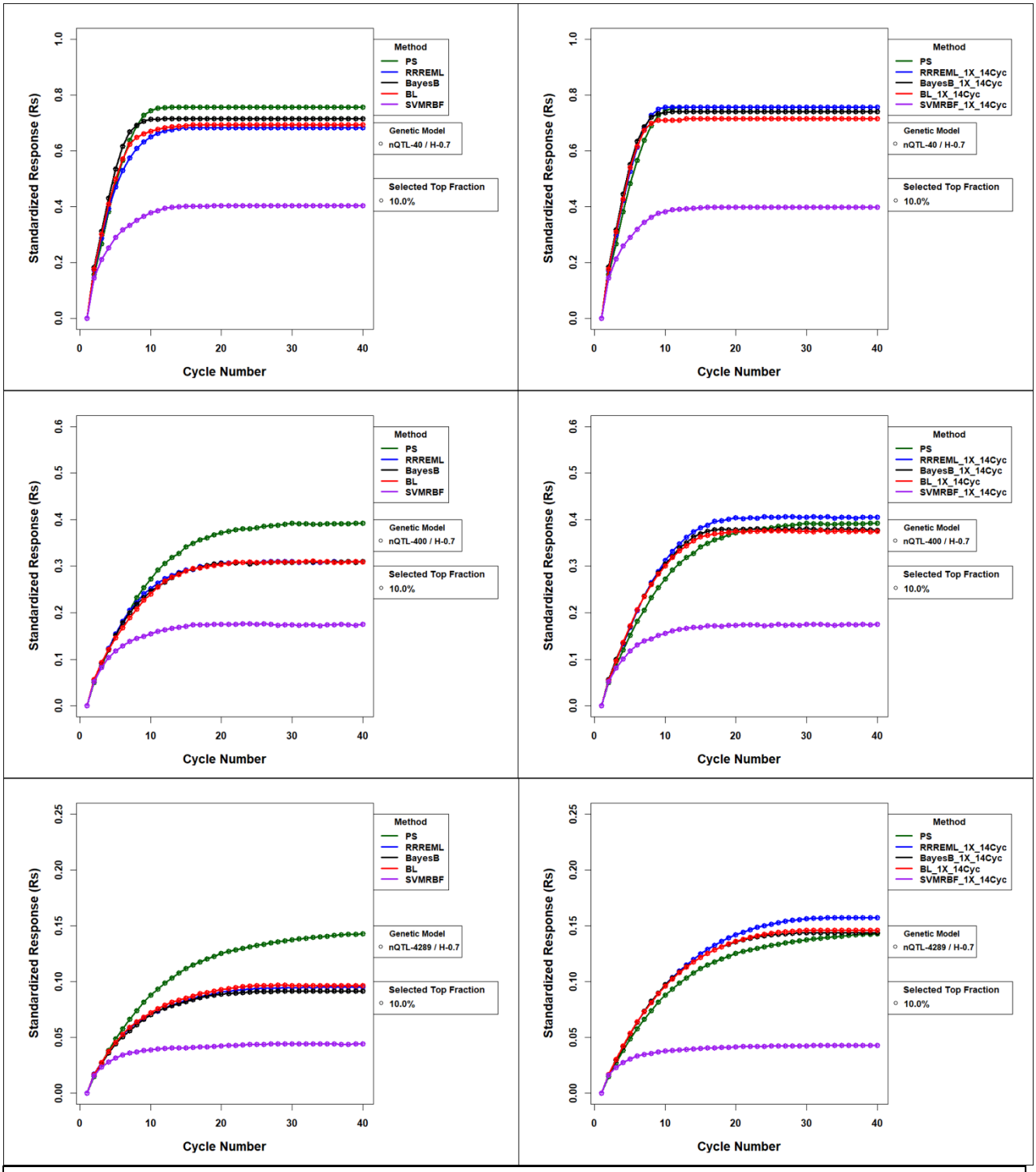

Figure 5 Standardized Responses for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction Forty cycles of standardized responses to selection of 10% of 2000 soybean RILs per cycle. Standardized responses are plotted by selection methods without (left panels) and with (right panels) model updating using prior cycles as training sets for the four genotypic prediction models. Phenotypic selection (PS) is not updated and hence is the same in the left and right panels. The top panels consist of responses for genetic architectures consisting of 40 simulated QTL. Middle panels consist of responses for genetic architectures consisting of 400 simulated QTL and the bottom panels

consist of responses for genetic architectures consisting of 4289 simulated QTL. All 40, 400, and 4289 simulated QTL are responsible for 70% of phenotypic variability in the initial population. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

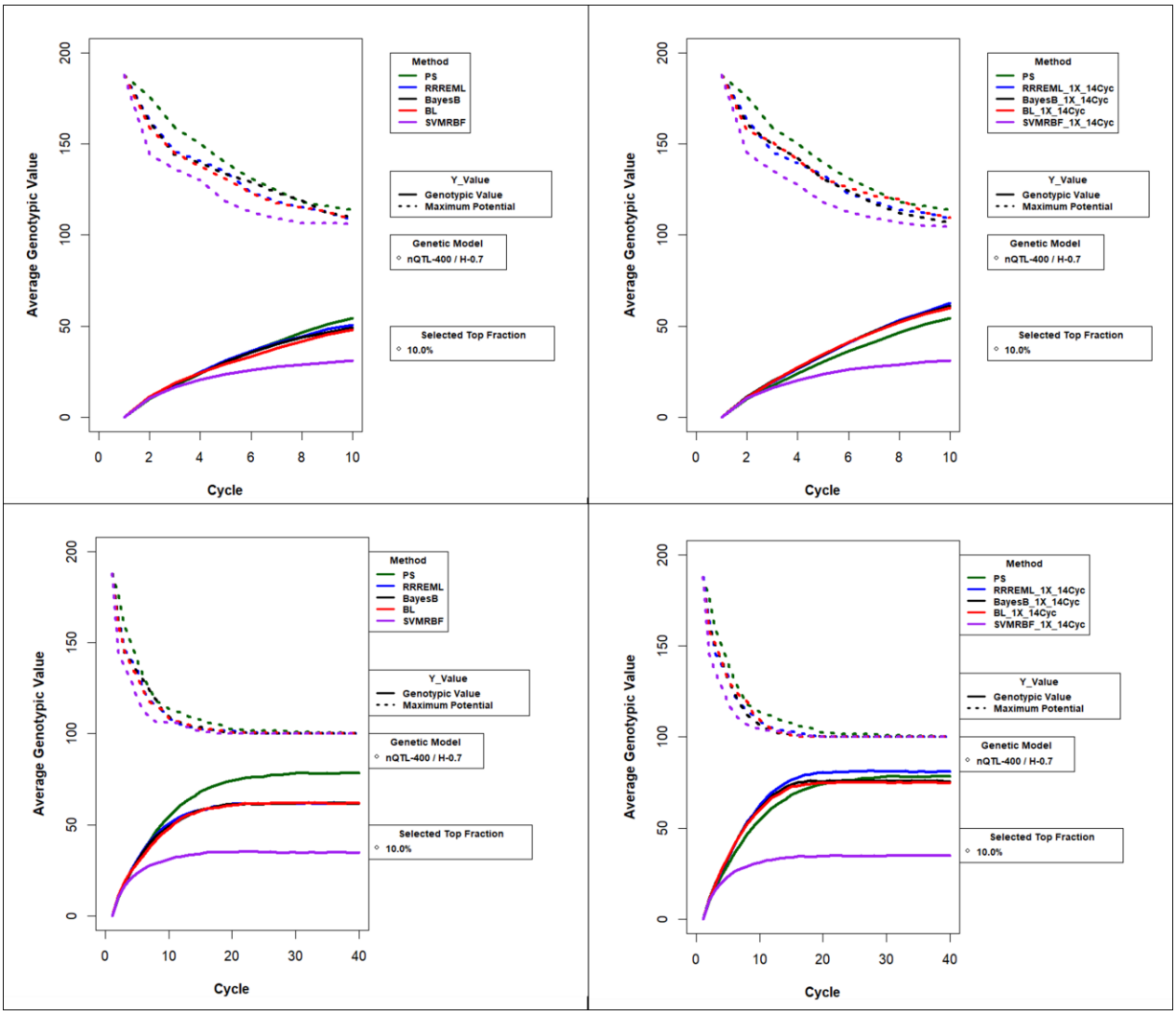

Figure 6 Average Genotypic Value and Maximum Genetic Potential for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction
Average genotypic value and maximum possible genotypic value in recurrent selection of 10% of 2000 soybean RILs per cycle. The values are plotted by selection methods without (left panels) and with (right panels) model updating using prior cycles as training sets for the four genotypic prediction models. Plots demonstrate decrease in maximum possible genotypic value due to loss of favorable alleles and increase in average genotypic value for 10 cycles (upper panel) and 40 cycles (lower panel) of selection. Phenotypic selection (PS) is not updated and hence is the same in the left and right panels. The top and bottom panels represent genotypic values for genetic architectures consisting of 400 simulated QTL responsible for 70% of phenotypic variability in the initial population. PS – Phenotypic Selection, RR-REML-Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF-Support Vector Machine with Radial Basis Kernel.

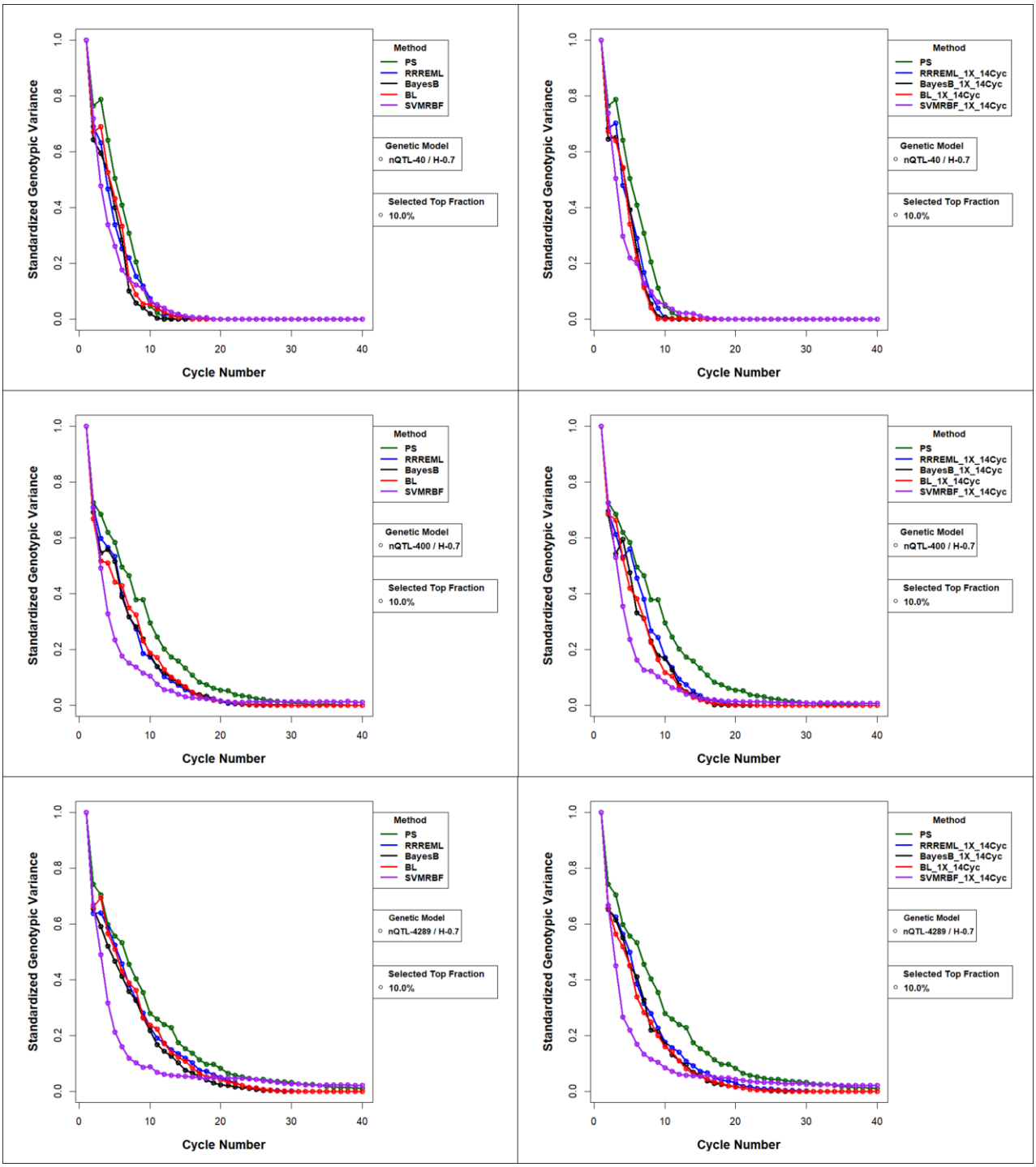

Figure 7 Standardized Genotypic Variance (Sgv) for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction Standardized genotypic variance without training set updating (left panels) and with training set updating using prior cycle training data (right panels) for the four GP models. PS has no updating and hence is the

same in both left and right panels. A) Training data from up to 14 prior cycles for 40 simulated QTL (top), 400 simulated QTL (middle) and 4289 simulated QTL (bottom) responsible for 70% of phenotypic variability in the initial population and top 10% of RILs with the greatest predicted values. PS – Phenotypic Selection, RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

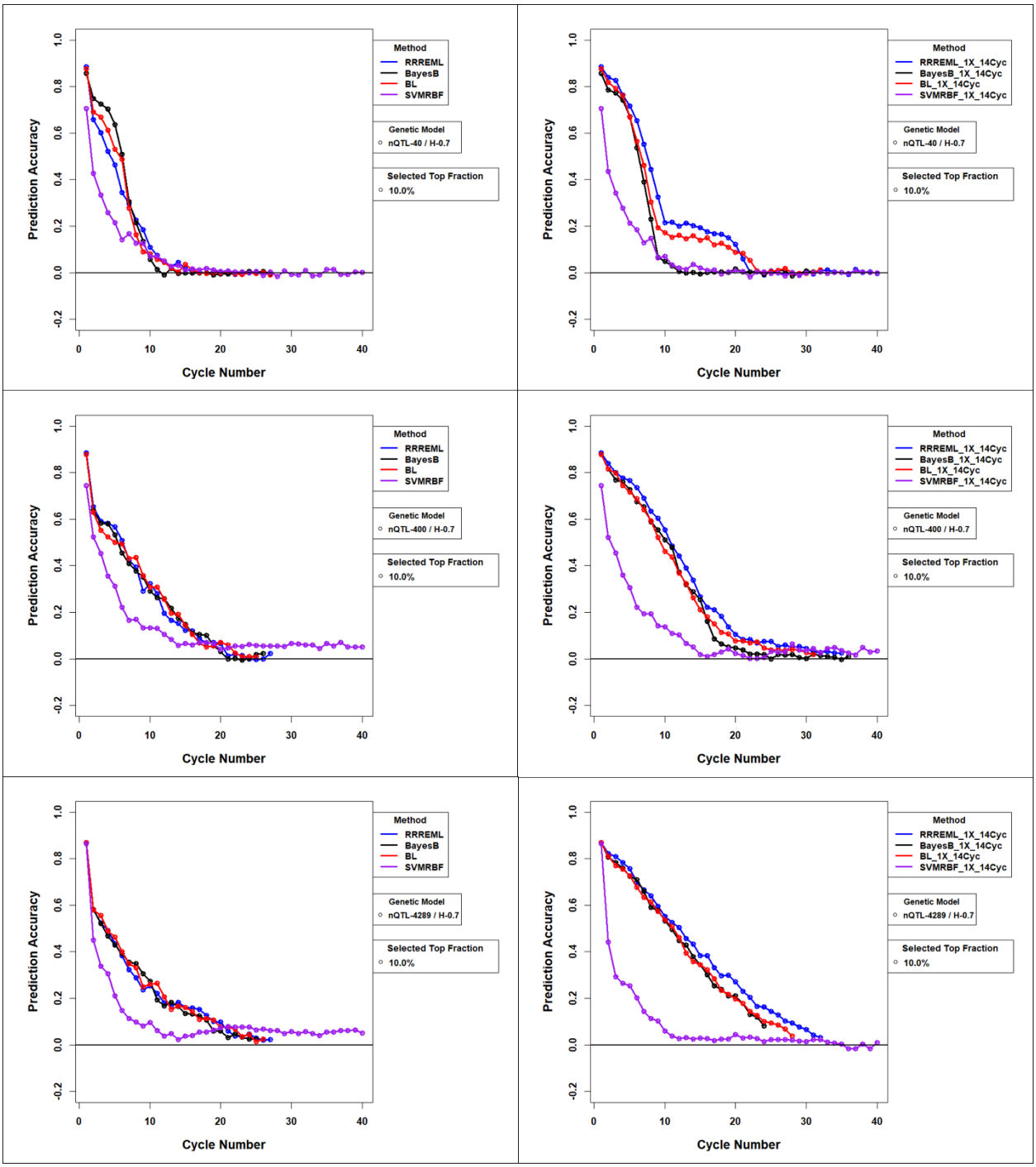

Figure 8 Estimated Prediction Accuracies for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction: Estimated prediction accuracies with updates to the training sets used in genomic prediction (GP) models. Training data from up to 14 prior selection cycles were used to update all four GP models for 40 QTL (top), 400 QTL (middle) and 4289 QTL (bottom) responsible for 70% of phenotypic variability in the initial population and top 10% of RILs with the greatest predicted values. RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

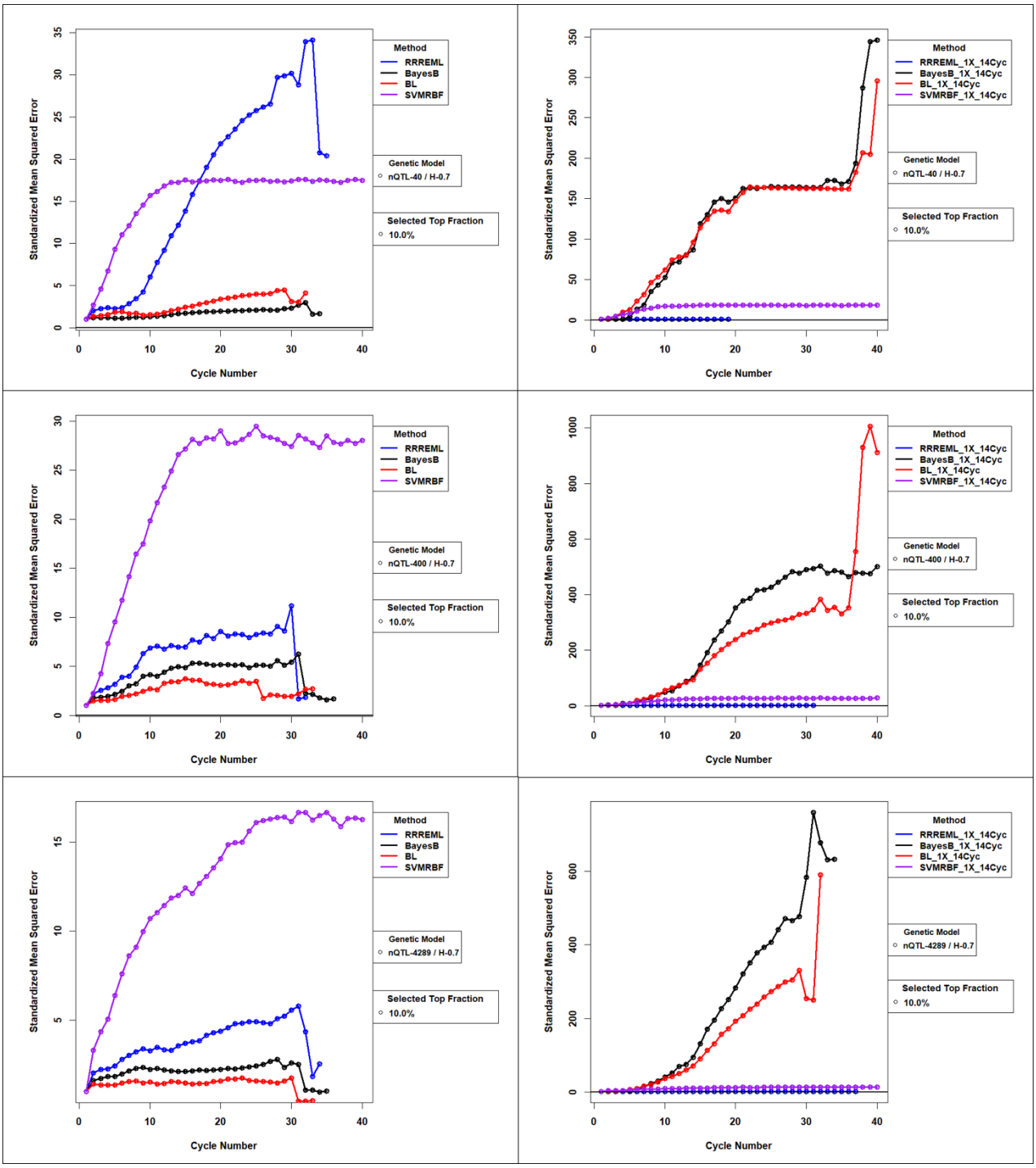

Figure 9 Standardized Mean Squared Error for Comparison of GS methods with and without Updating for 0.7 H and Top 10% Selected Fraction: Mean Squared Error of GP models with updates to the training sets used in genomic prediction (GP) models. Standardized MSE (>=1) is estimated as the ratio of MSE for GP models in cycle 'c' to MSE for GP models trained with founder population of RILs. While MSE for RRREML model were lesser with model updating, MSE for bayesian methods increased orders of magnitude in late cycles of selection with

model updating. SVMEBF showed relatively constant MSE with and without updating. Training data from up to 14 prior selection cycles were used to update all four GP models for 40 QTL (top), 400 QTL (middle) and 4289 QTL (bottom) responsible for 70% of phenotypic variability in the initial population and top 10% of RILs with the greatest predicted values. RR-REML- Ridge Regression with Restricted Maximum Likelihood, BL – Bayes LASSO, and SVMRBF- Support Vector Machine with Radial Basis Kernel.

Table 1: Factorial Design

Factors	Number of Levels	Values for Levels
Number of QTL	3	40, 400, 4289
Heritability	2	0.7, 0.3
Selection Intensity	3	2.67, 2.34, 1.75
Selection Method	5	i) PS- Phenotypic value ii) GS – GP(RR-REML) iii) GS - GP(Bayes B) iv) GS - GP(Bayes LASSO) v) GS – GP(SVM, Radial basis Function Kernel)
Model Update: number of prior cycles used in training sets	4	i) 0 previous cycles ii) 10 previous cycles iii) 12 previous cycles iv) 14 previous cycles
Total Number of unique combinations	360	
Total Number of Simulations	3600 with 10 reps /condition	

Table 2 Packages in R for Parametric and Non-Parametric Models with Tuning Parameters

GS Model	Package (R)	Model Tuning Parameters (BGLR package)
Ridge Regression	REML-EM (custom R script Xavier 2019)	EM algorithm for estimation of parameters with REML method without using matrix inversion
Bayesian LASSO	BGLR (Perez et al. 2014)	Priors for varE (df=3,S=0.25); varU (df=3,S=0.63); lambda(shape=0.53,rate=5e-5) type='random',value=30), nIter=20000, burnIn=2000, thin=1
Bayes B	BGLR (Perez et al. 2014)	nIter=41000, burnIn =1000, df0=4, R2=0.7
SVM	Rgtsvm (Wang et al. 2017)	SVM with Radial basis function kernel on GPU

Table 3 Tradeoff Table for GS Methods with/without Model Updates Tradeoff table to support decision for selecting the best method to achieve possible objectives including maximum gain in 5, 10, 20, 30 and 40 cycles of recurrent selection. The methods are ranked for each of the objectives based on percent gain in genetic response relative to phenotypic selection on non-isolated families using a hub network mating design PS with no update. Two sets of objective weights are provided to define the cumulative relative importance of the objectives: (A) the weighted rank of methods are estimated with emphasis only on the first 10 cycles and no emphasis on the remaining cycles, (B) the weighted rank of methods are estimated with equal emphasis on all forty cycles.

	Cumulative Relative Weights for Short-term Objectives	**0.5**	**1**	**0**	**0**	**0**
		Rs (Rank of method) in Cycle				
		5	**10**	**20**	**30**	**40**
Weighted Rank	**Selection Method - Training Set**					
1	**RR-REML-14**	0.17 (3)	0.31 (1)	0.4 (1)	0.4 (1)	0.41 (1)
2	**BayesB-14**	0.17 (2)	0.3 (2)	0.38 (2)	0.38 (4)	0.38 (4)
3	**BL-14**	0.17 (1)	0.3 (3)	0.38 (3)	0.38 (5)	0.37 (5)
4	**RR-REML-0**	0.15 (4)	0.25 (6)	0.31 (6)	0.31 (8)	0.31 (8)
5	**BL-0**	0.15 (8)	0.24 (8)	0.3 (8)	0.31 (8)	0.31 (8)

	Cumulative Relative Weights for Short-term and Long-term Objectives	**0.125**	**0.25**	**0.5**	**0.75**	**1**
		Rs (Rank of method) in Cycle				
		5	**10**	**20**	**30**	**40**
Weighted Rank	**Selection Method - Training Set**					
1	**RR-REML-14**	0.17 (3)	0.31 (1)	0.4 (1)	0.4 (1)	0.41 (1)
2	**BayesB-14**	0.17 (2)	0.3 (2)	0.38 (2)	0.38 (4)	0.38 (4)
3	**BL-14**	0.17 (1)	0.3 (3)	0.38 (3)	0.38 (5)	0.37 (5)
4	**PS**	0.15 (7)	0.27 (5)	0.37 (5)	0.39 (3)	0.39 (3)
5	**RR-REML-0**	0.15 (4)	0.25 (6)	0.31 (6)	0.31 (8)	0.31 (8)

CHAPTER 3. STRATEGIES TO ASSURE OPTIMAL TRADE-OFFS AMONG COMPETING OBJECTIVES FOR GENETIC IMPROVEMENT OF SOYBEAN

Vishnu Ramasubramanian[1, 2], William Beavis[1]
[1]George F. Sprague Population Genetics Group, Department of Agronomy,
[2]Bioinformatics and Computational Biology Graduate Program
Iowa State University, Ames, Iowa,
US - 50010

Modified from a manuscript under review in Frontiers in Genetics

Abstract

Plant breeding is a decision making discipline based on understanding project objectives. Genetic improvement projects can have two competing objectives: maximize rate of genetic improvement and minimize loss of useful genetic variance. For commercial plant breeders competition in the marketplace forces greater emphasis on maximizing immediate genetic improvements. In contrast public plant breeders have an opportunity, perhaps an obligation, to place greater emphasis on minimizing loss of useful genetic variance while realizing genetic improvements. Considerable research indicates that short term genetic gains from Genomic Selection (GS) are much greater than Phenotypic Selection (PS), while PS provides better long term genetic gains because PS retains useful genetic diversity during the early cycles of selection. With limited resources must a soybean breeder choose between the two extreme responses provided by GS or PS? Or is it possible to develop novel breeding strategies that will provide a desirable compromise between the competing objectives? To address these questions, we decomposed breeding strategies into decisions about selection methods, mating designs and whether the breeding population should be organized as family islands. For breeding populations organized into islands decisions about possible migration rules among family islands

were included. From among 60 possible strategies, genetic improvement is maximized for the first five to ten cycles using GS, a hub network mating design in breeding populations organized as fully connected family islands and migration rules allowing exchange of two lines among islands every other cycle of selection. If the objectives are to maximize both short-term and long-term gains, then the best compromise strategy is similar except a genomic mating design, instead of a hub networked mating design, is used. This strategy also resulted in realizing the greatest proportion of genetic potential of the founder populations. Weighted genomic selection applied to both non-isolated and island populations also resulted in realization of the greatest proportion of genetic potential of the founders, but required more cycles than the best compromise strategy.

1. Background

Historical responses to selection of commodity crops has been enabled by decreasing the number of years between cycles of recurrent selection, by increasing the number of replicable genotypes (selection intensity) and by increasing the number of field trials (heritability on an entry mean basis). In other words, genotypic improvements from responses to selection in commodity crops over the last 50 years (Specht et al. 2014) required monetary investments that became part of the exponential rise in seed costs during the same time (Byrum et al. 2017; USDA-ERS 2020). Since the emergence and adoption of Genomic Selection (GS), it has been possible to increase the numbers of genotypes that are evaluated, i.e., selection intensity, without significant increases in numbers of field plots (Bernardo 2007, 2008; Asoro et al. 2011; Heslot et al. 2012; Nakaya and Isobe 2012; Emily and Bernardo 2013; Crossa et al. 2014; Beyene et al. 2015; Bassi et al. 2016; Marulanda et al. 2016; Jonas and de Koning 2016; Hickey et al. 2017; Goiffon et al. 2017).

In a companion manuscript we reported an investigation of various factors on response metrics to recurrent selection of soybean lines derived from founders of the SoyNAM population (Ramasubramanian and Beavis 2020). The combinatorial set of factors consisted of phenotypic selection (PS) and four commonly used GS methods, training sets (TS), selection intensity (SI), number of QTL (nQTL), and heritability (H) on an entry mean basis. While interactions among all factors affected all response metrics, only the impacts of GS methods, SI and TS are factors that plant breeders can control. All GS methods provided greater responses than PS for at least five cycles, but PS provided better responses to selection as response from GS methods reached a limit. These results are consistent with reports by Goddard (2009), Jannink (2010) and Liu et al. (2015) demonstrated that the full genotypic potential of the founders is eliminated more quickly with GS than PS. In terms of factors that a soybean breeder can control, we found that SI's of 1.75 and use of Ridge Regression Genomic Prediction (RRGP) models updated every cycle with training data from all prior cycles of selection provided rapid response in the early cycles of selection and retention of genetic diversity for continued response to selection in later cycles, but we noted that further improvements might be made if the populations were organized into islands and mating designs other than the hub network were employed (Ramasubramanian and Beavis 2020). Herein we investigate strategies that soybean breeders can employ to find optimal trade-offs between maximizing genetic gain from selection and retaining useful genetic diversity.

The challenge of realizing genetic gains from selection and retaining useful genetic diversity in closed populations has been of interest since it was demonstrated that there are theoretical limits for response to selection in closed populations (Hill and Robertson 1968; Bulmer 1971). Trade-offs among objectives don't prohibit finding optima as long as optimality is

defined as a compromise among competing objective functions (Deb 2003; Konak et al. 2006; Shoval et al. 2012; Sheftel et al. 2013; Saeki et al. 2014).

Before the development of GS, quantitative geneticists working on domestic animal systems utilized mathematical programming modeling and operations research (OR) approaches to find near-optimal solutions to the challenge of assuring genetic gain and minimizing inbreeding per cycle of selection (Wray and Goddard 1994). The first publication using OR approaches to address multiple objectives in plant breeding was applied to selection of multiple traits (Johnson et al. 1988). Generally OR approaches involve three activities: 1) define objectives using measurable metrics, 2) translate the objectives into mathematical programming models consisting of objective functions, decision variables and constraints, 3) find an algorithm that will provide values for the decision variables resulting in optimal solutions to the mathematical programming model (Rardin 2017).

If the genetic improvement project wants to assure genetic gain and retain useful genetic diversity then there are two competing objectives for which a trade-off needs to be optimized. This represents an example of a multi-objective optimization (MOO) problem (Deb 2003, 2011; Rardin 2017). After translating each of the objectives into an objective function, there are several strategies for finding the optimal solution (Deb 2003). The two most commonly used strategies are known as the ε -constraint and the weighted sum. The ε -constraint method consists of identifying one of the objectives, e.g., maximize genetic gain, and translate other objectives, such as minimize inbreeding, into decision variables that can be constrained in a linear, integer or quadratic mathematical programming model (Haimes 1971). In other words, translate the MOO mathematical model into a Single Objective Optimization (SOO) model for which there exist computational algorithms capable of finding the optimum solution (Frank and

Wolfe 1956; McCarl et al. 1977; Lazimy 1982). Framing the ε -constraint method requires definition of metrics for genetic diversity or inbreeding. In animal breeding this method became known as Optimum Contribution Selection (OCS: Wray and Goddard 1994; Brisbane and Gibson 1995; Meuwissen 1997; Grundy et al. 1998; Meuwissen et al. 2001). Subsequent to development of GS, OCS was modified to maximize Genomic Estimated Breeding Values (GEBVs) and the realized relationship matrix was used to constrain inbreeding in what became known as Genomic OCS (GOCS) (Sonesson et al. 2010; Schierenbeck et al. 2011; Woolliams et al. 2015).

The second well-established approach to a MOO challenge is known as the weighted sum method. The weighted sum method assigns weights, $\omega_i \in [0, 1]$ and $\sum \omega_i = 1$, to each of the 'i' objective functions and an algorithm is employed to find the values for the decision variables that minimize all objective functions simultaneously (Zadeh 1963). Breeders will recognize the weighted sum method as a selection index composed of weighted parameters for genetic gain and inbreeding, or equivalently genetic diversity. If genomic information is available, GEBV's can be used to maximize genetic gain and the realized relationship matrix can be used to minimize inbreeding resulting in a genomic selection index (GSI) that can be calculated for all genotypes. (Carvalheiro et al. 2010; Clark et al. 2013).

Both ε-constraint and weighted sum methods are referred to as preference methods (Deb 2003) where the constraints or relative weights have been predetermined. For defined preferences there exist exact optimization algorithms if Karush-Kuhn-Tucker (KKT) conditions are met (Karush 1939; Kuhn and Tucker 1951). An exact optimization solution guarantees that no other feasible solution will be a better solution for the specified set of constraints or weights. Unfortunately, it is difficult to predetermine these values because they require forecasting the

relative economic values of genetic gains and retention of useful genetic diversity. For commercial plant breeding projects competition in the marketplace will force much greater emphasis on maximizing genetic gains than retaining genetic diversity. In contrast public soybean breeders have an opportunity, perhaps an obligation, to retain useful genetic diversity while realizing genetic gains for quantitative traits of agronomic importance. Because each plant breeding project has unique relative trade-offs, evolutionary algorithms (EAs) have been adopted to provide multiple solutions on an efficient (Pareto) frontier of solutions to competing objectives (Deb 2003, 2011; Konak et al. 2006). Decision makers then decide which of the solutions have the appropriate relative emphasis on the competing objectives.

Genetic algorithms (GAs) are a class of EAs that are based on recurrent selection of breeding populations and were developed to find computational solutions to large combinatorial problems (Goldberg 1989; Luque 2011). In a canonical GA, selected solutions are pooled together into a set of solutions. Subsequently the individual solutions are randomly sampled for pairwise "matings" to create a new set of solutions for evaluation and selection. The algorithm is iterated until there are no improvements in the sets of solutions. Computational analogs of mutation or recombination, are utilized to move from local optima to global optima. A subclass of GAs, known as parallel GAs maintain structure among subsets of individual solutions and enable the subsets to independently find different solutions for different domains (Luque 2011). The parallel GA system is analogous to the concept of genetic subpopulations (Falconer and Mackay Figure 3.2, 1996). Island Model GAs allow for exchange of solutions among subpopulations. Island model GAs (IMGAs) are distinct from canonical GAs in terms of properties and behavior because evolution happens locally, within island, and globally, among islands. Island model parameters consist of number of islands, island size, selection pressure

within each islands, numbers of migrants, migration frequency, connectedness or topology of islands and emigration and immigration policies among islands (Whitley 1999; Skolicki 2007 a, b).

Rather than investigate the trade-off between objective functions, Jannink (2010) proposed that it would be possible to retain useful genetic diversity in GS by weighting low frequency alleles with favorable estimated genetic effects. Simulations with Weighted Genomic Selection (WGS) resulted in greater responses across 24 selection cycles of recurrent selection than unweighted GS, using RRBLUP values, for both low and high heritability traits. However, the initial rates of response using WGS were less than responses from application of PS and less than GS. The response using WGS was better than response from PS after twenty cycles of selection, but the responses relative to GS depended on the number of simulated QTL and heritability. Decay of LD between marker and QTL is one of the factors that can slow responses using GS relative to PS (Hickey et al. 2014; Xavier et al. 2016), although decay of LD did not contribute to responses in the initial cycles using WGS. The rate of inbreeding per cycle is also greater with GS than with PS, whereas it is similar to PS when WGS is applied. The rate of fixation of favorable alleles is lower for WGS than GS resulting in larger numbers of cycles of genetic improvement before response to selection reaches a limit (Jannink 2010). Efforts to balance the response in early cycles and later cycles have included addition of parameters to WGS (Sun and Van Raden, 2014) and dynamic weighting of rare alleles depending on the time horizon for the breeding program (Liu et al. 2015). Low frequency favorable alleles are given greater weights, drawn from a Beta distribution, in initial cycles, and the weights tend towards unity as the number of cycles of selection approaches a predefined time horizon. This shifts the balance towards retaining greater genetic variance in earlier cycles.

The applications of GS, GOCS, GSI and WGS assume that selected individuals will be randomly mated. Typically, plant breeders do not randomly mate selected genotypes, rather most use selected genotypes that exhibit the most desirable selection metrics, e.g., GEBVs, to serve as "hub" parents in networked crossing designs (Guo et al. 2013; Guo et al. 2014). Such Hub-Network (HN) mating designs (MD's) apply greater weights to genetic contributions from hub genotypes resulting in amplified loss of genetic diversity relative to random mating by reducing the effective population size.

As soybean breeders have become aware of the potential impacts due to loss of genetic diversity from use of GS, they have used various *ad hoc* methods to avoid crosses between related genotypes (Diers, Graef, Lorenz, Cianzio, Singh, personal communications). After quantitative geneticists working on animal breeding systems demonstrated that it is possible to use the GSI strategy with an EA to identify optimal pairs of mates (Kinghorn 2011; Pryce et al. 2012; Woolliams et al. 2015), plant quantitative geneticists developed and investigated various versions of GSI and GOCS for plant breeding (Akdemir and Sanchez 2016; Lin et al. 2017; Cowling et al. 2017; Beukelaer et al. 2017; Gorjanc et al. 2018; Allier et al. 2019). Notice that the computational demand to find the optimum on the non-decreasing efficiency frontier created by all possible constraint values or relative weights in all NC_2 mating pairs is particularly well suited for application of EA's. Also, it should be noted that Akdemir and Sanchez (2016) referred to their implementation of GOCS as efficient GS. Last, we note that optimal mate selection has been referred to as optimal cross selection in plant breeding applications (Gorjanc et al. 2018; Allier et al. 2019), unfortunately with the same acronym as OCS. To distinguish optimal cross selection from optimal contribution selection, we do not use an acronym for optimal cross selection.

In addition to evaluating traditional PS, GS, and GOCS, Akdemir and Sanchez (2016) proposed and evaluated a novel mathematical programming model, referred to as genomic mating (GM). They formulated the problem as minimizing a linear function of inbreeding plus a negative risk function for the realized relationship matrix of N_p possible parents. Inbreeding is a function of the expected genetic diversity among N_c progeny from the N_p parents and is weighted by a parameter that controls allelic diversity among all N_p parents. Risk is determined for each cross as the sum of the expected breeding values of the progeny plus the expected standard deviations of marker loci weighted by a parameter that controls allelic heterozygosity of the relative contributions of the marker loci to the GEBVs. Thus, risk is similar to the usefulness criterion defined by Schnell (1983 as cited in Melchinger et al. 1988) of a selected proportion of the population and the weighting parameter reflects the breeders' emphasis on its importance. They demonstrated that their GM formulation is equivalent to an optimization problem of minimizing inbreeding subject to defined level of risk, denoted □. The solution needs to calculate risk and inbreeding for the range of acceptable □ values for N_c progeny from N_p parents, i.e., $(N_pC_2)^{Nc} / N_c!$ (Akdemir and Sanchez 2016) developed a Tabu-search GA to determine the efficiency frontier between inbreeding and risk. In an updated version, (Akdemir 2018) used a GA to find the complete set of non-dominated solutions (Deb 2003, 2011) that comprise the efficiency frontier for the three criteria of Gain (G), Inbreeding (I) and Usefulness (U) values in the objective function. This allows selection of a subset of solutions for evaluation obviating the need for conducting a grid search across all possible values.

Akdemir and Sanchez (2016) demonstrated the utility of their genomic mating approach using simulations of recurrent selection beginning with two founders for a trait composed of simple additive genetic architecture. The QTL were evenly distributed across a simulated

genome consisting of three diploid linkage groups. Their results indicated that the efficiency frontier can be selected to produce responses across 20 cycles that were better than PS and as good as GS and GOCS for the first five to seven cycles and better than PS, GS and GOCS thereafter (Akdemir and Sanchez 2016). They did not include WGS for comparison in their study.

Recognizing that IMGA's are very efficient at finding global optima Yabe *et al* (2016) suggested that computational island models could be used to create efficient and effective breeding plans for plant breeders. Even though computational IMGA's allow the software developer to change mutation and recombination rates, which are not under the control of plant breeders, structures of breeding populations based on island models could offset loss of useful genetic variability through regulation of exchange of genotypes among sub-populations. It is not unusual for plant breeders of crops that are easily self-pollinated to routinely evaluate, select and recurrently cross lines derived from one or two specific bi-parental crosses. In the vernacular of commercial soybean and maize breeders this is known as "working a population". Yabe et al (2016) demonstrated GS on populations organized as islands of families provided greater response to selection than GS after the seventh of 20 cycles of RGS. Their founder population consisted of lines derived from *in silico* crosses of six homozygous rice lines with an elite rice variety. They isolated the six families of RILs for recurrent selection using GS with no or occasional exchange of selected lines among the family islands. While their results appeared to be similar to WGS, they did not compare their results with WGS. They also suggested that the trade-off between genetic gain and retention of useful genetic variance could be improved by adjusting the number and frequencies of migrants among sub-populations.

Inspired by Akdemir and Sanchez (2016) and Yabe et al (2016), we hypothesized that a breeding strategy that organized the breeding populations as island families and utilized a genomic mating MD would provide small soybean genetic improvement projects with the ability to minimize the trade-offs between maximizing genetic gain and minimizing loss of useful genetic variability. Within the IM organized populations we evaluated three migration policies among the families. For both non-isolated and island models we applied three selection methods and four mating designs. To evaluate the potential of these combinations of methods to realize genetic gains while retaining useful genetic diversity, we compare outcomes from simulated recurrent selection applied to contemporary soybean germplasm adapted to MZ II and III using a set of metrics (Ramasubramanian and Beavis 2020). The metrics include the standardized genotypic value (R_s), the most positive genotypic value (*Mgv*) among RILs selected in cycle c, the standardized genotypic variance (*Sgv*), the average expected heterozygosity (Hs), the lost genetic potential of populations based on the number of favorable alleles that are lost.

2. Methods

2.1 Simulations.

Initial sets of soybean lines were generated by simulating crosses of 20 contemporary homozygous lines representing diversity of soybean germplasm adapted to MZ's II and III with IA3023 to generate *in silico* F1 progeny (Ramasubramanian and Beavis 2020). Individual F1's from each of the 20 crosses were self-pollinated *in silico* for four generations to generate 100 lines per family forming populations of 2000 lines organized into 20 families with genotypic information at 4289 genetic loci (Song et al. 2017). Thus the genetic structure of the initial simulated populations is similar to that used in the experimental SoyNAM investigation (Guo et

al. 2010; Takuno et al. 2012; Song et al. 2015; Song et al. 2017; Xavier A et al. 2017; Diers B et al. 2018).

As reported previously (Ramasubramanian and Beavis 2020), there were 3818 polymorphic loci in the combined population of 20 families and an average of 773 polymorphic loci within each of the families for the initial founding sets of lines. The variance among families was ~ 34 polymorphic loci. Across the 20 families of cycle 0 (C0) lines, average expected heterozygosity was 0.09 with an estimated variance of $4.4*10^{-7}$ among families. The average estimated G_{st} value across the genome for the initial founding set of RILs was 0.32, as determined by the 'diff_stats' function in the mmod R package (Jombart 2008; Ryman and Leimar 2009; Jombart and Ahmed 2011; Ramasubramanian and Beavis 2020). Average pairwise 'Fst' estimated using 'pairwise.fst' in 'hierfstat' R package (Goudet 2005) among the 20 families in simulated SoyNAM data is 0.20. Pairwise 'Fst' is a measure of population differentiation among pairs of populations, which is estimated as the ratio of difference between the average of the expected heterozygosity of the two populations and total expected heterozygosity of the pooled populations to total expected heterozygosity of the pooled populations. Whereas the average Fst using genotypic data from SoyNAM project among 40 families is 0.09 with a maximum pairwise Fst of 0.15 and a minimum Fst of 0.007 (Ramasubramanian and Beavis 2020).

2.2 Combinations of Factors.

We evaluated 60 combinations of factors (Table 1) that could influence responses to recurrently selected populations derived from a set of founder genomes representing the diversity of contemporary soybean germplasm adapted to MZ II and III in North America (Mikel et al. 2010; Diers et al. 2018). The treatment factors included structure of breeding populations,

selection method, and mating design. The structure of the breeding populations included organizing 20 sub-populations representing the original 20 founder families, referred to as family islands (FI), and populations in which the family structures were not retained after the initial founder population was created, referred to as non-island (NI) populations.

Previously we demonstrated that development of homozygous lines for phenotypic evaluation will limit the numbers of segregating linkage blocks with effective QTL effects each cycle of selection (Ramasubramanian and Beavis 2020). Consequently, we chose to designate only 400 polymorphic marker loci as simulated QTL. The QTL were distributed uniformly among the SNP loci and each contributed equal additive effects of ± 0.5 units to the total genotypic value of a line. Thus, cycle C0 lines derived from the founders had an average genotypic value of zero and the potential to create genotypic values ranging from -200 to +200. Phenotypic values were simulated by adding non-genetic variance sampled from an N (0, σ) distribution to the simulated genotypic values, where σ was determined by the heritability. Herein we report only simulated broad sense heritability values on an entry mean basis of 0.7. The non-genetic variance was held constant across subsequent cycles of selection. Thus, heritability is expected to decline with every cycle of selection due to the loss of additive genetic variance.

Phenotypic selection (PS), genomic selection (GS) and weighted genomic selection (WGS) were applied recurrently to both population structures. Recurrent selection applied to the non-island populations consisted of ranking all lines in a given cycle (Figure 1) according to the selection metric and retaining 10% for crossing to create the next cycle of lines. In terms of standardized selection differential, this corresponds to a selection intensity, 1.75. For selection of lines organized into FI's, 10% of the lines are selected within FI's (Figure 2). Subsequently, 20%

of lines might be migrants from other FI's depending on migration rules (Table 1). Metrics used for selection include simulated phenotypic values for PS, genome estimated breeding values (GEBVs) for GS and weighted genome estimated breeding values for WGS. We used the weighting function used by Jannink (2010) for estimating weighted genome estimated breeding values. The weighting functions are provided in Supplementary Table 1. Previous results indicated that among GS methods, Ridge Regression (RR) provided the best compromise between short term and long term responses (Ramasubramanian and Beavis 2020), thus we only used RR to train GP models for GS. RR was implemented with a method that employs Expectation Maximization to obtain Restricted Maximum Likelihood Estimates of marker effects (Xavier 2019).

For both GS and WGS the training models were updated every cycle of selection with data sets from all prior cycles. Since average within family prediction accuracies are lesser than prediction accuracies from a combined TS comprising of RILs from all the families (Ramasubramanian and Beavis 2020), we used a combined TS comprising of RILs from all the families. Training sets for each cycle were obtained by randomly sampling 1600 lines from the set of 2000 lines for each cycle. The most accurate predictions and maximum genetic responses were obtained with training data that is cumulatively added every cycle. For purposes of this manuscript, model updating refers to retraining the model with data from the current cycle as well as all prior cycles that were cumulatively added.

Subsequent to selection, four mating designs were applied to create the next cycle of lines (Table 1). To simulate theoretical truncation selection, selected lines were randomly mated (RM). The chain rule (CR), a.k.a., a single round-robin mating design (Yabe et al. 2016), is an alternative to RM that assures all selected lines contribute to the subsequent cycle of evaluation

and selection. In contrast to the attempt to assure equal representation of selected lines through RM and CR, most soybean breeders use a mating design that assures most progeny will be derived from crosses of a few lines that exhibit the most desirable performance (Guo et al. 2013; Guo et al. 2014). Because the metaphor of hubs with spokes represents the preference for crossing most selected lines to a few "hub" lines, we refer to this mating design as a hub network (HN) and is the mating design used in our previous investigation (Ramasubramanian and Beavis 2020). The fourth mating design, genomic mating (GM), uses mathematical objective functions to assure that defined breeding objectives are used to identify pairs of crosses from among the selected lines. Genomic Mating was implemented with the 'Genomic Mating' R package (Akdemir et al. 2018). As originally described GM combines selection and mating in a single step, but we decomposed the steps to provide comparable outcomes from all other combinations of selection methods, mating designs and organized populations. One of the implications is that all the selected lines are crossed in the GM method as in the CR method, however the contributions are optimized to minimize rate of inbreeding while maximizing gain.

2.2.1 *Genomic mating in non-isolated families*. In a selected set of 200 lines there are 200C2 (19900) combinations of parental pairs. To solve the objective function w.r.t an initial population of parental pairs, 250 initial populations of 200 combinations of parental pairs are sampled from 19900 combinations (19900C200) for the GA algorithm to solve.

2.2.2 *Genomic mating in populations organized as family islands*. In island selection, ten lines are selected from each of the 20 family islands. Within each island 45 (10C2) combinations of parental pairs are possible (Supplementary Figure1). To solve the objective function w.r.t an initial population of parental pairs, 250 initial populations of 10 combinations of parental pairs are sampled with replacement to keep the population size equal to the NI populations for the GA

algorithm. For each of the 20 families, the GA algorithm is applied to the initial subset of 250 out of all possible combinations (45C10). The other parameters for the GA algorithm are the same for both NI and FI populations. The GA algorithm selects non-dominated elite solutions (Deb 2003, 2011) and mates of non-dominated elite solutions for 50 iterations with a mutation probability of 0.8 (Supplementary Figure1). Examples of pseudocode are provided in (Akdemir and Sanchez 2016) and the Genomic Mating R package manual (2018). It is important to note that the parameters values in the GA algorithm can be optimized and the set of solutions in the pareto-front can be explored for better solutions by other methods such as NSGA-II, NSGA-III, SPEA-1, SPEA-2 and other recent improved versions of GA for better convergence rate and quality of solutions, determined by the proximity to global optimum (Deb 2011; Seada and Deb 2018) (Supplementary Figure1).

2.3 Migration Rules among Family Islands.

In addition to applying selection methods and mating designs to both population structures, there are many possible rules that affect migration among islands. A preliminary investigation of migration rules that was implemented included: 1) Frequency of migration - never, once every two cycles and every cycle of recurrent selection. 2) The proportion (10% and 20%) of immigrants that will be included in crosses responsible for creating the next cycle of lines. 3) Migration can be either in one direction or it can be reciprocal among family islands. Based on the preliminary investigation, we decided to set migration rules bi-directional migration between both immigrant and emigrant islands of two lines once every other cycle of selection.

2.3.1 *Migration Policies among family islands* included three levels included “Best Island” (BI), “Random Best” (RB), and “Fully Connected” (FC). Migration policy (MP) refers to the nature of island topology specifying connections between emigrant and immigrant islands.

For the BI policy, emigrant lines are selected from the island with most desirable genotypic value and the selected lines can emigrate to no more than 10 islands. Given a bi-directional migration rule, the emigrant island also receives two immigrants from the islands that received the emigrants. For a RB policy, an emigrant island is selected randomly from a set of islands with high genotypic values, while the migration pattern itself is similar to BI policy. For the FC policy, every island is connected to every other island and lines migrate from emigrant islands with high values to randomly selected immigrant islands (Supplementary Figure2).

Note that migration factors are irrelevant for populations that did not maintain the structure of FIs and they are irrelevant for FI's that do not experience migration. Thus the treatment design is not a complete factorial, rather the complete set is comprised of responses for 60 combinations of factors.

2.4 Modeled Response to Recurrent Selection.

The averaged genotypic value for each cycle, c, of recurrent selection was modeled with a linear first order recurrence equation:

$$f_0(\mathrm{c})y_{(\mathrm{c}+1)} + f_1(\mathrm{c})y_{(\mathrm{c})} = g(\mathrm{c}) \qquad \text{(Eqn 1)}$$

Where c is a sequence of integers from 0 to 39 representing each cycle of recurrent selection from cycle 1 to 40 and f_0, f_1 and g are constant functions of c. By rearranging the equation we note that the response in cycle c+1 can be represented as

$$y_{(\mathrm{c}+1)} = -\frac{f_1(\mathrm{c})}{f_0(\mathrm{c})}y_{(\mathrm{c})} + \frac{g(\mathrm{c})}{f_0(\mathrm{c})} \qquad \text{(Eqn 2)}$$

Since the ratios $f_1(\mathrm{c})/f_0(\mathrm{c})$ and $g(\mathrm{c})/f_0(\mathrm{c})$ are constants, we can represent the response in cycle c+1 as

$$y_{(\mathrm{c}+1)} = \alpha y_{(\mathrm{c})} + \beta \qquad \text{(Eqn 3)}$$

If y_0 specifies the average genotypic value of the first cycle of RILs derived from the founders, then (3) has a unique solution (Goldberg 1958; Ramasubramanian and Beavis 2020):

$$y_c = \alpha^c y_0 + \beta\frac{1-\alpha^c}{1-\alpha} \text{ if } \alpha \neq 1 \qquad \text{(Eqn 4)}$$
$$y_c = \alpha^c y_0 + \beta c \qquad \text{if } \alpha = 1$$

An alternative representation of (eqn 4) for the situation of $\alpha \neq 1$ is

$$y_c = \alpha^c (y_0 - y') + y' \qquad \text{(Eqn 5)}$$
$$\textit{with } y' = \frac{\beta}{1-\alpha},$$

, where α is less than 1 for genotypic response to recurrent selection and *y*' represents the asymptotic limit to selection (Goldberg 1958; Ramasubramanian and Beavis 2020). An illustration of the values of the sequence of c=0 to 39 for a range of α and β values can be found in our previous study (Ramasubramanian and Beavis 2020). The model derived curves can be interpreted as response to selection as a function of the frequencies of alleles with additive selective advantage, selection intensity, time and effective population size (Robertson 1960). The parameters, y_o, α, and β, were estimated with a non-linear mixed effects method implemented in 'nlme' functions in the 'nlme' and 'nlshelper' packages (Pinheiro and Bates 2000; Baty et al. 2015; Pinheiro et al. 2019).

Since the limits of responses are approached asymptotically, the number of cycles required to reach half of the limits before there is no longer response to selection is referred to as the half-life of the recurrent selection process is referred to as the half-life (Robertson 1960; Dempfle 1974; Kang 1979; Cockerham & Burrows 1980; Kang and Namkoong 1980; Kang 1987; Kang and Nienstaedt 1987). From the first order recurrence equation (5), the half-life is estimated as

$$t_{1/2} = \ln(0.5) \ / \ \ln(\alpha) \qquad \text{(Eqn 6)}$$

, when y0 is ‘0’ and the asymptotic limit is estimated as y' (Ramasubramanian and Beavis 2020).

2.5 Analyses of variance (ANOVA) of Modeled Response to Recurrent Selection.

ANOVA is used to evaluate the impact of factors and their interactions on the modeled responses to global and island recurrent selection. The analyses of variance used single level nlme models with modeled (eqn 5) responses grouped by combinations of treatment factors. We analyzed the variance among modeled responses using AIC, BIC and Likelihood metrics that were grouped based on combinations of treatment variables consisting of population type, selection method, mating design, and migration policy for a constant level of migration frequency, migration size and migration direction for one genetic model consisting of 400 simulated QTL responsible for 0.7 H with equal additive effects (Table 1). For a discussion of ANOVA using non-linear mixed effects models refer (Pinheiro et al 2000; Zuur 2009; Baty, et al. 2015; Pinheiro et al. 2019; Oddi et al. 2019; Ramasubramanian and Beavis 2020).

In the first phase of model fitting, we fit a random intercept model for estimating both alpha and beta in the recurrence equation using the ‘nlme’ R package. Estimates of modeled parameters from nlsList models were retained as starting values for fixed effects. Multiple ANOVA of ‘nlme’ objects representing the models were used to identify combinations of factors with significant effects on the non-linear response. The model with the lowest AIC score was selected as the best model. The best random intercept model in the first phase of model fitting process M15 and models with combinations of three factors (M11-M14) showed auto-correlation of residuals. Since auto-correlation violates the independence assumption, the correlation among

residuals was modeled using AR-1 correlation structure. Since the genotypic values across cycles in recurrent selection are correlated, fitting AR-1 correlation doesn't remove the correlation unless cycles are used as co-variates. However, using cycles as a co-variate makes the model fitting very time consuming and often has larger AIC scores than models without cycles as co-variates. The Model M15 with AR-1 correlation structure was further refined by modeling variance components using 'varIdent' structure in 'nlme'. The process of fitting, selecting and refining mixed effects models is similar to our previous study (Ramasubramanian and Beavis 2020) and is described in the vignettes in R package 'SoyNAMSelectionMethods'.

2.6 Evaluations of Responses to Recurrent Selection

Evaluations were conducted on both modeled and genotypic values using a set of metrics described in (Ramasubramanian and Beavis 2020) and defined below. The estimated population half-life and asymptotic limits used the estimated parameters, α and β of the first order recurrence model. The average genotypic values were used to estimate the standardized genotypic value (Rs) and maximal genotypic value (Mgv). Maximum possible genotypic potential of the founders provided a reference for number of favorable alleles retained in the population. The loss of genotypic potential is characterized by reduction in the standardized variance of genotypic values (Sgv) and estimated heterozygosity (Hs). In addition, efficiency of conversion of loss in genotypic variance into genetic gain (Rs_var) provides a way to assess gain in genotypic value and loss of genetic variance simultaneously. In island model selection, the different impacts of selection strategies on the genotypic variance at island or global levels are assessed using intra-island Sgv, inter-island and global variance of genotypic values. A schematic diagram of the factors and evaluation metrics used to characterize the responses to recurrent selection is provided in Figure 3.

2.6.1 *Evaluation Metrics*. The standardized genotypic value, R_s (Meuwissen et al. 2001; Liu et al. 2015; Ramasubramanian and Beavis 2020), was estimated every cycle of selection as the proportion of maximum genotypic potential (200 units) relative to the average genotypic value of 2000 lines in C0 (eqn 7). Values range from 0-1 with the value of 1 corresponding to the maximum possible genotypic value with the genetic model and 0 corresponding to the average genotypic value of C_0.

$$R_s = \frac{R_c}{(R_m - R_0)} \qquad \text{(Eqn 7)}$$

R_s - Standardized genotypic value
R_0 - Average genotypic value of RILs produced by founders
R_c – (Average genotypic value in cycle 'c' – R_0)
R_m - Maximum possible genotypic value (200)

Since we previously evaluated genetic improvement of soybean using PS and the HN mating design in NI populations, we used PS with a selection intensity of 1.75 for NI population and HN mating design (designated as NI-PS-HN) as a reference for comparing other selection and mating designs. A standardized relative genotypic response, ΔRs_c is calculated in equation (8) as the percentage of the difference in standardized genotypic values, Rs_c, in each cycle c.

$$\text{Percent Gain in } Rs_{c\,(\text{Design-x})} = \frac{Rs_{c\,(\text{Design-x})} - Rs_{c(\text{NI-PS-HN})}}{Rs_{c(\text{NI-PS-HN})}} * 100 \qquad \text{(Eqn 8)}$$

$Rs_{c\,(\text{Design-x})}$ - standardized response for Design - x in cycle 'c'
$Rs_{c\,(\text{NI-PS-HN})}$ - standardized ersponse for NI - PS - HN design in cycle 'c'

The standardized genotypic variance (*Sgv*) defined as the estimated genotypic variance divided by the estimated genotypic variance of the initial sample of lines from C0 was used to evaluate the changes in estimated genotypic variance across cycles of recurrent selection. Note that values for the *Sgv* range from zero to one.

Efficiency of genetic improvement is a metric used to estimate the proportion of genetic improvement that was obtained through loss of genetic diversity from recurrent selection (Gorjanc et al. 2018). Efficiency is estimated as the slope in linear regression in linear regions of response curves. However, responses to recurrent selection in the absence of mutation are inherently non-linear (Robertson 1960; Hill and Robertson 1968; Bulmer 1976; Ramasubramanian and Beavis 2020). For purposes of evaluating the relative contribution of lost genetic variance to genetic response in both linear and non-linear segments of the response curve, we introduce the standardized genotypic variance of the response, Rs_Var, calculated with equation (9).

$$Rs\_var = \frac{G_c - G_0}{SdG_0 - SdG_c} \qquad \text{(Eqn 9)}$$

G_c-average genotypic value of the set of RILs evaluated in cycle c
G_0-average genotypic value of the founding set of RILs
SdG_0 - estimated standard deviation from genotypic values of founding set of RILs
SdG_c - estimated standard deviation from genotypic values of RILs in cycle c

The numerator term represents difference in average genotypic values of a population in cycle ‘c’ from cycle ‘0’ normalized to standard deviation of genotypic values in cycle ‘0’. The denominator represents difference of standard deviation of genotypic values between cycles ‘0’ and cycle ’c’ normalized to the standard deviation of genotypic values in cycle 0 (Ramasubramanian and Beavis 2020). For the NI populations, Rs_Var was estimated by calculating the variance of simulated genotypic values. Standardizing the estimated genotypic variance with respect to the maximum genotypic values in the initial population, results in values

that range from 0-1. For the FI populations, the genotypic variances can be split into within and between island genotypic variance. The three measures we used to estimate the global diversity of populations, inter-island diversity and within island diversity are provided in the documentation of the R package.

3. Analyses and Data Availability

Simulated data and software codes are available as part of the R package 'SoyNAMSelectionMethods' (Supplementary File1). Documentations for downloading and using the package are available at http://gfspopgen.agron.iastate.edu/SoyNAMSelectionMethods_v2_2020.html. The SoyNAM founder genotypic and phenotypic data are available in SoyBase (Grant et al. 2010).

4. Results

4.1 Rates and Limits of Responses to Recurrent Selection.

Factors common to NI populations and FI populations such as mating design and selection method as well as factors specific to discrete and island model selection had significant effect on estimated population half-lives and asymptotic limits. Half-lives for selection methods on NI populations ranged from 3.83 to 16.10 cycles with a mean of 9.62 cycles and asymptotic limits ranged from 71.64 to 160.76 with a mean of 115.97 (58% of the maximum possible potential in the founders). Compared to NI populations, half-lives for discrete selection (DS) methods were very low ranging from 1.97 to 2.89 cycles with a mean of 2.43 cycles and asymptotic limits ranged from 28.42 to 38.30 and a mean of 33.12 (16.5% of the maximum possible potential in the founders) (Supplementary File2; Supplementary Figure3). Estimated half-lives for island model selection methods were greater, on the average than NI methods ranging from 4.24 – 32.04 cycles with a mean 13.45 cycles and asymptotic limits ranged from

47.54 to 198.82 with a mean of 116.8 (58.5 % of the maximum possible potential in the founders) (Supplementary File2; Supplementary Figure3).

4.2 ANOVA of Modeled Genotypic Values.

There is strong evidence from the analyses of variance (Supplementary File3) that the modeled genotypic values across cycles of selection depend on interactions among selection method, mating design and migration policy. The most parsimonious model included all combinations of factors indicating interactions among all factors have statistically significant influences on recurrent responses to selection and requires unique estimates of α, and β in (3) for each of the combinations of factors (M15 in Supplementary File3). For all combinations of factors, we report only migration involving bi-directional migration of two migrants every other cycle. Among the factors that affect only FI populations, migration frequency had significant effects on rate and the asymptotic limits for response to selection, whereas migration direction and size had relatively small effects on rates and no significant effect on the asymptotic limits for response to selection. Rates and genotypic values at the limits of response for a given selection method and mating design also depend on genetic architecture and heritability (data available on request). Rather than belabor the specific outcomes from all possible combinations of factors that affected the modeled responses, the remainder of the reported results are restricted to results from simulations with 400 QTL responsible for 70% of phenotypic variability.

4.3 Responses to Recurrent Selection of Non-Isolated Lines.

There were 12 combinations of selection methods and mating designs that were applied to lines of NI populations. The greatest genotypic values (Rs) were attained with WGS (Figure 4 and Supplementary Figure 4). Genomic selection using RRBLUP values resulted in greater responses than PS in early cycles while WGS produced greater responses than PS in later cycles (Figure 4; Supplementary Figure4). Weighted genomic selection followed by the CR mating

design resulted in the greatest realization of genetic potential before reaching a limit. Genomic selection using RRBLUP values followed by a hub network (HN) mating design resulted in the greatest rates of response in the first ten cycles and if followed by RM, provided the greatest responses in the first 20 cycles. When the GM design is applied to selected lines to obtain specified crosses according to optimization criteria, the responses in the first 15 cycles were larger than obtained with RM, whereas responses after the 20th cycle were less than responses for other mating designs (Figure 4 and Supplementary Figure 4).

The responses measured as maximum genotypic values (Mgvs) produced response patterns similar to Rs. Use of WGS followed by the chain rule (CR) mating design resulted in an average Mgv of 125 (62.5% of the maximum potential in the founders) followed by PS and GS using RRBLUP values in the 40th cycle. Genomic selection followed by the HN mating design (NI-GS-HN) realized greater Mgvs relative to other combinations of factors only in the early cycles (Supplementary Figure 5).

The rate at which maximum genotypic potential decreased across cycles of selection was reflected in the estimated number of lost favorable alleles. Among the selection methods, GS using RRBLUP values lost genetic potential faster than PS and WGS. Among the mating designs, HN resulted in the fastest loss of genetic potential while RM lost genetic potential slower than any of the other mating designs. Genomic mating lost genetic potential at a rate that was intermediate between RM and HN mating designs. The CR design lost favorable alleles at rates that were similar to GM after GS, whereas after applying CR after PS and WGS, the loss of alleles was similar to RM (Figure 4).

The rates at which favorable alleles were lost exhibited similar patterns as the changes in standardized genotypic variance (Sgv) and expected heterozygosity (Hs) (Figure 4 and

Supplementary Figure 6). The application of RM and CR mating designs after selection helped maintain genotypic variance and heterozygosity for use in later cycles of recurrent selection. The HN mating design resulted in the fastest loss of Sgv and Hs (heterozygosity) while the GM design demonstrated losses of Sgv and heterozygosity that were intermediate between HN and RM/CR designs.

Rates of inbreeding are larger for GS compared to PS and WGS in the first 10-15 cycles. The RM and CR mating designs demonstrated the slowest rates of inbreeding, whereas inbreeding with the GM and HN mating had high rates of inbreeding before responses to selection became limited (Supplementary Figure 7 & 8). The estimates of genotypic responses, standardized to genotypic variance (Rs_Var), were the greatest in the first 20-30 cycles with CR, RM and GM mating designs while the HN mating design lost the greatest amount of phenotypic variance after GS, PS and WGS (Supplementary Figure 9 & 10).

4.4 Responses to Recurrent Selection of Lines Organized as Family Islands.

The genotypic values when the population reached a limit using Discrete Selection (DS), where there is no exchange of lines between islands, were as much as 67% less than the values when limits were reached in the non-isolated counterpart populations (Supplementary Figure 11). Among the DS methods, GS and WGS with GM design (designated DS-GS-GM and DS-WGS-GM) provided the greatest genotypic values at the response limits. Between 10-15% of the maximum potential in the founder populations were realized within the first 10-15 cycles with DS (Supplementary Figure11). Mgvs followed a pattern similar to Rs, and Sgvs mirrored the response pattern in DS (Supplementary Figure 11).

In contrast to isolated family islands (FIs) with DS, genotypic values at the limits to selection responses were larger using BI, RB and FC migration policies among islands. Among

the selection methods applied to the FI populations, GS and WGS realized the greatest genetic potential before reaching limits of responses. The impacts of mating designs on the responses to selection applied to FI populations are distinct from mating designs in NI populations. In the NI populations RM and CR mating designs provided the greatest genotypic values before response to selection became limited, whereas in the FI populations GM provided the greatest genotypic values when coupled with BI and RB migration policies. The fully connected (FC) migration policy with the largest migration rates, produced responses that were similar among the HN, CR, RM and GM designs.

As noted above, the best responses to selection in the first 10 to 20 cycles of NI populations were obtained using GS followed with a HN or GM mating design (respectively designated NI-GS-HN and NI-GS-GM in Figure 4). The greatest short-term responses to selection in FI populations were obtained using either GS or WGS followed by the HN mating design coupled to a FC migration policy (IM-GS-HN-FC and IM-WGS-HN-FC in Figure 4 and Supplementary Figure 4). Only slightly slower rates in the first 10 to 20 cycles were obtained using GS and WGS followed by the GM design coupled to a FC migration policy.

Given a FC migration policy, the largest standardized genotypic responses at the limits to response (0.59 – 0.61) were obtained using GS or WGS with HN, CR, RM and GM designs Given a RB migration policy, GS and WGS followed by GM design produced the greatest realization of genetic potential before the 40^{th} cycle (0.59 -0.6) compared to (0.3-0.4) with HN, CR and RM designs (Figure 4 & Supplementary Figure 4). The BI policy showed a pattern similar to that of RB, but at a slower rate of response (Figure 4 & Supplementary Figure 4).

Mgvs followed a pattern similar to Rs for most of the island selection methods. In contrast to selection in NI populations where PS and WGS resulted in the greatest Mgvs in 20-40

cycles, GS in FI populations resulted in larger Mgvs (124.6) than island PS (119.9) by the 40th cycle.

Rates of decrease in maximum available potential is influenced by factors such as selection intensity, SM and MD. Relative to NI populations, island selection retains allelic diversity in the combined population as selection depletes variance only within islands and not across islands (Figure 6). Such loss in maximum potential is not always reflected in rates of responses. Relaxed selection intensity will result in retention of genetic variance with no significant increase in response as it is observed with BI and RB migration policies when combined with RM designs for PS, GS and WGS.

IM-GM-FC design showed the least rate of decrease of Hs values for PS, GS and WGS reflecting a greater potential retained in the population followed by IM-GM-RB and IM-GM-BI designs. IM-HN-BI and IM-HN-RB designs showed the most rapid decrease in Hs across 40 cycles of selection, whereas CR and RM designs with RB and BI migration policies showed intermediate rates of decrease of Hs. There is an oscillatory pattern in the decrease of Hs, where Hs increased with every migration event in early cycles. In late cycles, the magnitude of increase in Hs due to migration event decreased and the oscillatory pattern dampened to a continuous decrease as the populations approached the limits of responses (Supplementary Figure 6).

Island PS demonstrated lesser rates of inbreeding compared to island GS and WGS. RM design showed the least rates of inbreeding among the four MDs for BI, RB and FC migration policies (Supplementary Figure 7 and 8). CR design followed a pattern similar to HN or GM depending on the SM. FC migration policy demonstrated lesser rates of inbreeding compared to

BI and RB policies. BI policy demonstrated the largest rates of inbreeding. GM design demonstrated rates of inbreeding that were intermediate between RM and HN/CR designs (Supplementary Figure 8).

Rs_Var for island selection with FC migration policies were larger than that observed with NI populations, demonstrating larger efficiency of converting loss of genetic variance into gain. However, with FC policy, all MDs showed a similar pattern (Supplementary Figure 9). Whereas Rs_Var for island selection with BI and RB policies were comparable to that of global selection with PS and GS, except for GM design, which showed larger Rs_Var after 10-20 cycles of selection (Supplementary Figure 10).

4.4.1 *Diversity within and among islands* The average within island genotypic variance decreased towards zero through forty cycles of selection, whereas global and inter-island genotypic variance increased before becoming limited. The rates of decrease in average within island genotypic variance was influenced by SM, MD and MP. Both GS and WGS demonstrated similar patterns of lost genotypic variance within islands and rates of loss with both were faster than PS (Figure 5). The HN mating design demonstrated the fastest loss of within island genotypic variance followed by RM, CR and GM designs. The FC migration policy provided the slowest loss of within island genotypic variance followed by RB and BI migration policies (Figure 5). Notice, however, an oscillatory pattern in which within island genotypic variance increased with every migration event and decreased from selection in cycles where there were no migrants. For both the within island genotypic variance and the expected heterozygosity the magnitudes of oscillations dampened towards zero after 20-30 cycles of selection except for the GM mating designs coupled with BI and GM migration policies (designated IM-GM-BI and IM-

GM-RB respectively in Figure 5). The amplitude of increased genetic variance due to migration was greater for RB and BI migration policies with large spikes after 25-30 cycles of selection, while the amplitudes were smaller with the FC migration policies (Figure 5).

The largest values for inter-island genotypic variance were obtained with the RM mating design and BI and RB migration policies followed by CR and HN designs with BI and RB migration policies (Figure 7). Whereas, the FC migration policies demonstrated the smallest increases in inter-island genotypic variance through 40 cycles of selection (Figure 7). Recall that the FC migration policies provide the greatest migration rates among islands. Global genotypic variance in FI populations increased due to increase in inter-island genotypic variance. The BI migration policies demonstrated the largest global genetic variance for RM, HN and CR mating designs followed by the RB migration policies. Genomic mating under BI and RB migration policies provided intermediate rates of increasing global genotypic variance while the FC migration policy showed the least increase in global genotypic variance when coupled with the HN, CR, RM and GM mating designs (Figure 7).

Within the classes of migration policies, the migration frequency had significant influence on rates and limits of responses across most combinations of selection methods, mating designs and migration policies, while numbers of migrants significantly affected responses in only for a few combinations of factors. Both rates and limits of response decreased with fewer migrants for the HN mating design. For the RM design, exchange of migrants among FI's once in every three cycles provide the greatest genotypic values at the limits compared to responses with more frequent exchange. Migration size and migration direction had no significant effect on limits to selection responses (data available on request).

4.5 Tradeoffs between short-term and long-term gains from recurrent selection.

There were 12 combinations of selection methods and mating designs applied to NI populations and 48 combinations of selection methods, mating designs and migration policies applied to FI populations. From among the 60 methods, genomic selection using a ridge regression model followed by a hub model mating design in NI populations and weighted genomic selection followed by crosses using the chain rule in NI populations (respectively designated NI-GS-HN and NI-WGS-CR in Table 2 and Figure 4) demonstrated the greatest responses in the first 20 and last 20 cycles respectively. However, if the objective for genetic improvement is to maximize gain in the first 5, 10, 30 or 40 cycles, other combinations of the factors are needed to achieve the objective. If the breeding objective is to maximize rates of genetic improvement in five to ten cycles of recurrent selection then there are two best options: 1. Genomic selection using RRBLUP values followed by a hub model mating design in FI populations with fully connected migration policies, or 2. Genomic selection using RRBLUP values followed by a genomic mating design in FI populations with fully connected migration policies (respectively designated as IM-GS-HN-FC and IM-GS-GM-FC in Table 2). If the objectives are to maximize both short-term and long-term gains then the best solution was obtained by selecting with RRBLUP values followed by a genomic mating in FI populations and applying a fully connected migration policy (designated IM-GS-GM-FC in Table 2). Among the combinations applied on NI populations, weighted genomic selection followed by the CR mating design or RM resulted in largest long-term gains, while selection using RRBLUP values followed by a HN mating design provided the greatest short-term gains. Indeed, both GS and WGS demonstrated greater long-term responses than phenotypic selection in both NI and FI populations.

5. Discussion

5.1 Significance.

The challenge of finding optimal trade-offs among competing genetic improvement objectives has usually been approached by combining selection and crossing in a single step without consideration of population structure (Akdemir & Sanchez 2016; Beukelaer et al. 2017; Akdemir et al. 2019; Allier et al. 2019 a, b, 2020; Ramasubramanian and Beavis 2020). Akdemir & Sanchez (2016) combined selection and mating in their GM method. Beukelaer et al (2017) used weighted selection indices to maximize gain while retaining a threshold level of diversity. Among the three diversity measures they tested, indices that incorporate diversity measures to minimize loss of rare favorable alleles and minimize heterozygosity resulted in responses that were greater than WGS with truncation selection. Including diversity measures in a set offered advantage over truncation selection, as selected mate pairs retained rare favorable alleles better than WGS coupled with random mating. Allier et al (2019 a, b) included the impact of within-family selection to maximize genetic gain while minimizing loss of genetic variance, but they did not consider migration among families. And (Ramasubramanian and Beavis 2020) investigated GS methods for genetic improvement of soybean, but only considered the HN mating design applied to populations without regard to family affiliation. Herein we approached the challenge by disentangling breeding decisions into four distinct groups: 1) organization of the breeding population, 2) selection methods, 3) mating designs and 4) migration policies. Each of these were divided into possible alternatives within each group and treated as independent factors in orthogonal treatment combinations.

As with our previous investigation we found that the fastest rates of genetic improvement resulted when GS followed by the HN mating design is applied among all lines regardless of

their family affiliation. When combined, these three decisions have reinforcing effects on responses to selection. At the other extreme, when WGS is applied to populations organized as family islands followed by either CR or RM the tendency of all three to retain genetic diversity reinforce each other resulting in the largest genotypic values, but is not achieved until the later cycles of selection. Because the slopes of the curves resulting from WGS and PS at 40 cycles are still positive, it is possible that both selection methods could continue to produce greater genetic potential with more cycles of selection. In previous comparative studies, WGS produced long-term responses that are similar to methods such as Optimal Contribution Selection (OCS) and Expected Maximum – Haploid Value (EM-HPV) (Daetwyler et al. 2015; Muller et al. 2018). Herein when we applied WGS to lines regardless of family affiliation and followed selection by identifying optimal mate pairs using GM the genotypic values at the limits to response were greater than the genotypic values obtained with PS or GS followed by GM. This combination also retained the largest values for heterozygosity and favorable alleles across more cycles. However, the differences between responses to GS and WGS followed by GM were not significant when applied to lines organized into family islands.

Between the extreme response curves it was also possible to find many response curves with intermediate trade-offs between the objectives. For example applying WGS to lines that were not organized into islands followed by HN provided greater response rates than other combinations of factors involving WGS. Selection among lines organized into FIs resulted in responses that were larger or comparable to responses from NI populations for only a limited number of combinations of mating design (GM) and migration policies (RB and FC). This may be due to the small numbers of related lines on each island (20X smaller than the NI population). With such a small number selection can deplete all the genetic variance within the first 10 -15

cycles as demonstrated in discrete selection. When there is no migration, which is the major source of new genetic variability, the populations realized only 10-15 % of maximum potential in the founder populations even while optimizing for sustainable gain using the GM method. A relaxed selection intensity, where the top 20% of the lines in each island are selected can sustain responses for longer cycles as demonstrated in non-isolated and island selection with migration (Supplementary Figure 12).

As expected, even with small numbers of lines per island migration had a positive impact on the outcomes. It is known that intermediate levels of migration rate result in optimal tradeoffs between gain and diversity (Skolicki 2007 a, b: Obolski et al. 2017). However, the range of intermediate parameter values depend on the specific context and population genetic parameters. In our study, responses in FIs were larger than selection responses in NI only when migration events happened every cycle or once in two cycles. When migration event happened only once in 3 cycles of selection, the rates of responses in the early cycles were very low resulting in much lower genotypic values as responses to selection approached limits. Migration size and direction didn't have any significant impact on response within the small range of parameter values we tested for migration size and direction.

Also, we retained the best line within island during migration events and replaced the second best line in the ranked list of selected lines with the immigrant for the BI and RB policies. Whereas, for the FC policy, lines ranked from 2-6 are replaced. This replacement policy allows crossing between lines that are best within islands and immigrant lines from islands with high genotypic value resulting in high rates of response within islands. However, other policies that replace lines with low genotypic value with high genotypic values from immigrant islands will

reduce genetic diversity within islands and result in different outcomes compared to the policy we've implemented.

None-the-less, we found a very good tradeoff among the competing objectives. If a GS was applied to lines on FC islands and the selected lines were mated according to the pareto-optimal crosses identified using GM, then the combination preserved genetic variance for long-term gain with little penalty relative to the realized rates of improvement in early cycles by GS and the HN mating design. In summary, motivated by Akdemir and Sanchez (2016) and Yabe et al (2016), we demonstrate that it is possible to design breeding strategies to produce desired outcomes between the extremes of maximizing the rate of genetic improvement and maximizing the genetic potential of the population.

5.2 Interpretations.

The results can be interpreted from other perspectives to provide alternative insights. First genetic improvement can be viewed as single or multiple connected search strategies in genotypic space (Podlich and Cooper 1999; Cooper et al. 2002; Cooper et al. 2014). The single search strategy, a.k.a. global, corresponds to selection of lines in NI populations. The multiple connected search strategy, a.k.a. local, corresponds to selection of lines in multiple domains with infrequent exchange of lines. In addition to the perspective of global or local search strategies, selection can be viewed as cooperation vs. competition and exploitation vs. exploration. Thus, by tuning parameters that control relative levels of cooperation or exploration in global or local search strategies, it is possible to adjust the adaptive landscape.

Genotypes co-operate when they contribute to other genotypes' fitness values, whereas they compete when they reduce the fitness values of other genotypes thereby reducing their contribution to future generations. Intermediate levels of cooperation often accelerated shifts in adaptive peaks for bi-locus genetic models (Whitley 1999; Skolicki 2007; Obolski et al. 2017).

For global selection, selection methods and mating designs control contributions of genotypes within populations thereby controlling the level of cooperation. From this perspective, GS promotes competition, while PS and WGS emphasize cooperation on the adaptive landscape. By weighting rare favorable alleles, WGS promotes cooperation and effectively retains more of the genotypic potential of the founder's fitness landscape. Among mating designs, the HN used by most plant breeders, promotes competition over cooperation, whereas the CR and RM designs promote co-operation. The GM design provides a balance between cooperation and competition. For island selection methods, a FC migration policy provides the best balance between cooperation and competition. Further work is needed to identify optimal migration rules.

The concepts of exploitation and exploration are commonly used in EAs. In general, exploitation refers to processes such as selection that result in beneficial solutions, whereas exploration allows searches for solutions in new domains. In the breeding context, exploration maintains diversity (Goldberg 1989; Goldberg and Deb 1992; Whitley 1999; Skolicki 2007 a, b; Crepinsek et al. 2013). Because GS provides faster rates of genetic improvement than PS and WGS it is reasonable to interpret GS results as rapid exploitation, whereas PS and WGS allow exploration of new solutions, primarily through additional recombination opportunities. The HN mating design drives the populations to exploit resources for immediate gains, whereas CR and RM mating designs provide opportunities for the population to explore more of the fitness landscape. The GM mating design provides an opportunity to choose relative importance of exploitation and exploration. By treating population organization, selection methods and mating designs as orthogonal factors we were able to blur the boundary between exploitation and exploration with combinations of factors that mixed exploitation and exploration activities in distinct phases and simultaneously.

From the perspective of tradeoffs between exploration and exploitation, selection among and within islands enables exploration of diverse domains resulting in greater probabilities of finding solutions with greater genotypic values (higher peaks or limits at convergence), whereas global selection across a NI population of lines tends to get trapped in sub-optimal peaks or local optima (Cantu-Paz 2000; Skolicki 2007 a, b; Luque 2011; Crepinsek et al. 2013). In our study, this occurred with the GM design applied to the NI populations, where the crossing process within islands are optimized at a local level. In some implementations of island model selection, both the global and local states of islands are assessed every cycle of selection and a centralized global agent makes decisions to reach optimization objectives. We do not know whether such a bi-level optimization method will result in greater genotypic values before approaching limits to responses from selection.

5.3 Future Research.

By framing breeding strategies as combinations of population structure, selection methods, mating designs and migration policies we illustrated the potential of the approach for a small arbitrary soybean genetic improvement project. We did not consider the relative emphasis of objectives and constraints for any specific genetic improvement project. Consider first the structure of breeding populations. We compared a NI structure of lines with FI's created by individual crosses among the founders and then we selected within and among islands according to the same criteria. This might make sense within a single soybean genetic improvement project for lines adapted to MZ's II and III.

There are six public soybean genetic improvement projects adapted MZ's II and III. There are likewise about the same number of commercial soybean genetic improvement projects in the same MZs. All of these projects began at different times and were initiated with unique,

albeit overlapping, germplasm resources (Mikel et al. 2010). While all of the projects select lines with greater genotypic values for yield, the yield values are obtained from different, albeit overlapping environments.

From the perspective of soybean genetic improvement across regions within MZ's II and III, each genetic improvement project can be represented as an island where genotypes are exchanged among project islands based on annual evaluations in uniform regional trials and according to legal licensing rules. In practice project islands exchange only the best performing lines adapted to similar environmental conditions. None-the-less, soybean breeders will maintain useful genetic variability by exchanging lines among island projects. An advantage of island selection is that diversity among islands increases with selection, even when within island diversity decreases. Eventually, beyond 40 cycles of recurrent selection, genetic variability among islands will decrease as genetic variability among islands is lost to selection.

Future investigations of breeding strategies to optimize tradeoffs between rates of genetic gains and retention of useful genetic variance in soybean adapted to MZ's II and III should consider population structures within island projects that more accurately reflect those that currently exist. Also, future investigations should simulate genetic architectures with genotype x environment effects. It is well known that a line adapted to one environment may not perform well in other environments, and it is possible to define fitness values so that they include environmental effects. Third, future investigations should consider a broader set of migration rules and policies. The FC migration policy is considered the upper bound of island models as all islands are connected to every other island with maximum migration rates among islands. While our results indicate that this policy provided the results we don't think it will provide the best results for genetic architectures with genotype by environment interaction effects.

Fourth, we need to recognize islands in time because every cycle of selection discards useful genetic variability. A soybean germplasm resource project was set up (Mikel et al. 2010) to recover useful genetic variability lost during domestication of soybean (Nelson 2011). Rather than trying to build long bridges to islands located in the distant past, our results suggest that there should be a large amount of useful genetic variability that was discarded in the first few cycles of modern soybean breeding. For that matter, until response to selection reaches the half-life for the population, large amounts of useful genetic variability can probably be recovered from islands represented by recent cycles of discarded lines. These conjectures should be preceded by simulations to determine the potential benefit and costs associated with sampling lines in recently discarded islands.

Fifth, it should be clear that a predefined mating design does not take advantage of opportunities created by each cycle of progeny to optimize outcomes according to most project objectives. Thus, there continues to be a need for algorithms that efficiently and effectively identify crosses from among genotypes produced by each cycle of selection. It is tempting to adopt and investigate all EA strategies. However, only a subset are relevant to the practice of plant breeding (Hagan et al. 2012). For example, mutation and recombination rates can be controlled in a computational EA, whereas plant breeders cannot regulate these with current practices. None-the-less there are many opportunities for cross-disciplinary research between EAs and plant breeding. There is large body of literature concerning the properties of EAs and factors and strategies that affect convergence rates and quality of solutions (Goldberg 1989; Goldberg and Deb 1992; Whitley 1999; Skolicki 2007 a, b; Crepinsek et al. 2013; Obolski et al. 2017) and working with computational scientists should reveal novel methods to maximize the genetic potential of a breeding population in a minimum number of cycles.

Akdemir and Sanchez (2016) proposed only one of many possible GA's to identify pareto-optimal solution pairs. An approach introduced by Gaur and Deb (2016) and Mittal et al. (2020) would provide pareto-optimal solutions using statistical methods such as clustering and machine learning. The statistical knowledge can be used to improve the search for optimal solutions and establish several cycles of optimization. Conceptually, unveiling any relationship among pareto-optimal pairs in a genotypic space is likely to provide new knowledge regarding the characteristics of such complementary pairs. Also, by modeling responses with a first order recurrence equation or a non-linear mixed effects model to predict the half-life and asymptotic limits of selection have potential to improve the efficiency of genetic algorithms by providing repair operators to alter the trajectory of population evolution towards the desired optimal trade-offs.

Last, consider the challenge of stating explicit relative emphasis on objectives and definition of constraints for any specific genetic improvement project. As noted previously, this challenge exists because it requires assigning economic and agronomic value of short term genetic gains vs. the forecasted value of useful genetic variants that may be discarded each cycle of selection. As a thought experiment note that the trade-off objectives can be reduced to a single 'grand' objective of creating a genotype (line) with the genotypic value equal to the full genetic potential of the founders in a single cycle. For a genetic architecture consisting of two alleles at a single locus, achieving the single grand objective is trivial. Also it is possible to imagine that the grand objective can be achieved for a complex genetic architecture with infinite resources. Clearly, given genetic architectures of complex traits and resource constraints there are no feasible solutions to the grand objective, but it is a useful reference to serve as the goal.

Acknowledgments

Funding for this research was provided by the Department of Agronomy, Iowa State University, the North Central Soybean Research Program and an NSF grant (1830478). Supplemental funding for large scale computing was enabled by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation. XSEDE resources consisted of research allocations (DMS180041, DMS190003, DMS190015 & DMS190018) on PSC-Bridges Large Memory nodes for the simulations involving island model and genomic mating simulations. We want to thank Dr. Deniz Akdemir for discussions on implementing 'genomic mating' and Dr. Lizhi Wang for efficient programs to simulate meiosis. We also want to thank Dr. Alencar Xavier for sharing an efficient expectation maximization method for fitting ridge regression GP models.

References

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J: **Multi-objective optimized genomic breeding strategies for sustainable food improvement.** *Heredity* 2019, **122:**672-683.

Akdemir D, Sánchez JI: **Efficient Breeding by Genomic Mating.** *Frontiers in Genetics* 2016, **7**.

Akdemir D, Sanchez JI, Haikka H, Brum IB (2018) GenomicMating: Efficient Breeding by Genomic Mating. R package version 2.0.

Allier A, Lehermeier C, Charcosset A, Moreau L, Teyssèdre S: **Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection.** *Frontiers in Genetics* 2019, **10**.

Allier A, Moreau L, Charcosset A, Teyssèdre S, Lehermeier C: **Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression.** *G3: Genes|Genomes|Genetics* 2019, **9:**1469.

Allier A, Teyssèdre S, Lehermeier C, Moreau L, Charcosset A: **Optimized breeding strategies to harness Genetic Resources with different performance levels.** *bioRxiv* 2019**:**2019.2012.2020.885087.

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L: **Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats.** Iowa State University Digital Repository; 2011.

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J: **Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.).** *Plant Science* 2016, **242:**23-36.

Baty F, Ritz C, Charles S, Brutsche M, Flandrois J-P, Delignette-Muller M-L: **A Toolbox for Nonlinear Regression in R: The Package nlstools.** *Journal of Statistical Software; Vol 1, Issue 5 (2015)* 2015.

Bernardo R: **Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years.** *Crop Science* 2008, **48:**1649.

Bernardo R, Yu J: **Prospects for Genomewide Selection for Quantitative Traits in Maize.** *Crop Science* 2007, **47:**1082-1090.

Beukelaer HD, Badke Y, Fack V, Meyer GD: **Moving beyond managing realized genomic relationship in long-term genomic selection.(Author abstract).** *Genetics* 2017, **206:**1127-1138.

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, et al: **Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress.(RESEARCH)(Author abstract).** 2015, **55:**154.

Brisbane J, Gibson J: **Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions.** *International Journal of Plant Breeding Research* 1995, **91:**421-431.

Bulmer MG: **The effect of selection on genetic variability: a simulation study.** *Genet Res* 1976, **28:**101-117.

Byrum J, Beavis B, Davis C, Doonan G, Doubler T, Kaster V, Mowers R, Parry S: **Genetic Gain Performance Metric Accelerates Agricultural Productivity.** *Interfaces* 2017, **47:**442-453.

Cantú-Paz E: *Efficient and accurate parallel genetic algorithms / by Erick Cantú-Paz.* Boston, Mass.: Boston, Mass.: Kluwer Academic Publishers; 2000.

Carvalheiro R, Queiroz SAd, Kinghorn B: **Optimum contribution selection using differential evolution.** *R Bras Zootec* 2010, **39:**1429-1436.

Clark SA, Kinghorn BP, Hickey JM, van der Werf JHJ: **The effect of genomic information on optimal contribution selection in livestock breeding programs.** *Genet Sel Evol* 2013, **45:**44-44.

Combs E and Bernado R: **Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers.** *The Plant Genome* 2013, **6**.

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G: **Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction.** vol. 65. pp. 311-336. Melbourne; 2014:311-336.

Cooper M, Podlich D, Micallef K, Smith O, Jensen N, Chapman S, Kruger N: **Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops.** *Quantitative genetics, genomics and plant breeding'(Ed MS Kang) pp* 2002**:**143-166.

Cowling, W A, L Li, K H M Siddique, M Henryon, P Berg, R G Banks, and B P Kinghorn. "Evolving Gene Banks: Improving Diverse Populations of Crop and Exotic Germplasm with Optimal Contribution Selection." *Journal of Experimental Botany* 68, no. 8 (2016): 1927-39.

Črepinšek M, Liu S-H, Mernik M: **Exploration and exploitation in evolutionary algorithms: A survey.** *ACM Comput Surv* 2013, **45:**Article 35.

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, et al: **Genomic prediction in CIMMYT maize and wheat breeding programs.** *Heredity* 2014, **112:**48-60.

Daetwyler HD, Hayden MJ, Spangenberg GC, Hayes BJ: **Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity Relative to Genomic Selection.** *Genetics* 2015, **200:**1341.

Deb, Kalyanmoy. "Unveiling Innovative Design Principles by Means of Multiple Conflicting Objectives." *Engineering Optimization* 35, no. 5 (2003/10/01 2003): 445-70

Deb, Kalyanmoy. "Multi-Objective Optimisation Using Evolutionary Algorithms: An Introduction." In *Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing*, 3-34: Springer, 2011.

Deb K: *Innovization: Discovering Innovative Solution Principles Through Optimization.* Springer Publishing Company, Incorporated; 2014.

Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, et al: **Genetic Architecture of Soybean Yield and Agronomic Traits.** *G3: Genes|Genomes|Genetics* 2018, **8:**3367.

Falconer, D. S. *Introduction to Quantitative Genetics / D.S. Falconer and Trudy F.C. Mackay*. Edited by Trudy F. C. Mackay. 4th ed.. ed. Essex, England: Essex, England : Longman, 1996.

Frank M, Wolfe P: **An algorithm for quadratic programming.** *Naval research logistics quarterly* 1956, **3:**95-110.

Gaur A, Deb K: *Adaptive Use of Innovization Principles for a Faster Convergence of Evolutionary Multi-Objective Optimization Algorithms.* 2016.

Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, **136:**245-257.

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS: **Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection.** *Genetics* 2017, **206:**1675.

Goldberg DE: **Sizing populations for serial and parallel genetic algorithms.** In *Proceedings of the 3rd international conference on genetic algorithms*. Morgan Kaufmann Publishers Inc.; 1989: 70-79.

Goldberg DE: *Genetic algorithms in search, optimization, and machine learning / by David E. Goldberg.* Reading, Mass.: Reading, Mass. : Addison-Wesley Pub. Co.; 1989.

Goldberg DE, Deb K: **Massive multimodality, deception, and genetic algorithms.** *Urbana* 1992, **51:**61801.

Goldberg S: *Introduction to difference equations, with illustrative examples from economics, psychology, and sociology.* New York: New York, Wiley; 1958.

Gorjanc G, Gaynor RC, Hickey JM: **Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection.** *Theoretical and Applied Genetics* 2018, **131:**1953-1966.

Goudet J: **HIERFSTAT, a package for R to compute and test hierarchical F-statistics.** *Molecular Ecology Notes* 2005, **5:**184-186.

Grundy B, Villanueva B, Woolliams JA: **Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development.** *Genet Res* 1998, **72:**159-168.

Guo B, Wang D, Guo Z, Beavis WD: **Family-based association mapping in crop species.** *Theoretical and Applied Genetics* 2013, **126:**1419-1430.

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G: **The impact of population structure on genomic prediction in stratified populations.** *Theoretical and Applied Genetics* 2014, **127:**749-762.

Hagan S, Knowles J, Kell DB: **Exploiting Genomic Knowledge in Optimising Molecular Breeding Programmes: Algorithms from Evolutionary Computing (Evolutionary Computing for Molecular Breeding).** 2012, **7:**e48862.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Science* 2012, **52:**146-160.

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G: **Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation.** *Crop Science* 2014, **54:**1476-1488.

Hickey JM, Chiurugwi T, Mackay I, Powell W: **Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery.** *Nature genetics* 2017, **49:**1297.

Jannink J-L: **Dynamics of long-term genomic selection.** *Genetics Selection Evolution* 2010, **42:**35.

Johnson B, Gardner CO, Wrede KC: **Application of an Optimization Model to Multi-Trait Selection Programs.** *Crop science* 1988, **28:**723-728.

Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers.** *Bioinformatics* 2008, **24:**1403-1405.

Jombart T, Ahmed I: **adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.** *Bioinformatics* 2011, **27:**3070-3071.

Jonas E, de Koning DJ: **Goals and hurdles for a successful implementation of genomic selection in breeding programme for selected annual and perennial crops.** *Biotechnology & genetic engineering reviews* 2016, **32:**18.

Karush W: **Minima of functions of several variables with inequalities as side constraints.** *M Sc Dissertation Dept of Mathematics, Univ of Chicago* 1939.

Kinghorn BP: **An algorithm for efficient constrained mate selection.** *Genetics Selection Evolution* 2011, **43:**4.

Konak A, Coit DW, and Smith AE. "Multi-Objective Optimization Using Genetic Algorithms: A Tutorial." *Reliability engineering & system safety* 91, no. 9 (2006): 992-1007.

Kuhn H, Tucker A: **Nonlinear programming In Proceedings of 2nd Berkeley symposium (pp. 481–492).** *Berkeley: University of California Press[Google Scholar]* 1951.

Lazimy R: **Mixed-integer quadratic programming.** *Mathematical programming* 1982, **22:**332-349.

Liu H, Meuwissen TH, Sorensen AC, Berg P: **Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs.** *Genet Sel Evol* 2015, **47:**19.

Luque G: *Parallel Genetic Algorithms: Theory and Real World Applications.* Berlin, Heidelberg : Springer Berlin Heidelberg; 2011.

Marulanda J, Mi X, Melchinger A, Xu J-L, Würschum T, Longin C: **Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale.** *Theor Appl Genet* 2016, **129:**1901-1913.

McCarl BA, Moskowitz H, Furtan H: **Quadratic programming applications.** *Omega (Oxford)* 1977, **5:**43-55.

Melchinger AE, Schmidt W, Geiger HH: **Comparison of Testcrosses Produced from F2 and First Backcross Populations in Maize.** *Crop science* 1988, **28:**743-749.

Meuwissen TH: **Maximizing the response of selection with a predefined rate of inbreeding.** *Journal of animal science* 1997, **75:**934-940.

Meuwissen, T., B. Hayes, and M. Goddard. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157, no. 4 (2001): 1819-29.

Mikel MA, Diers BW, Nelson RL, Smith HH: **Genetic Diversity and Agronomic Improvement of North American Soybean Germplasm.** *Crop Science* 2010, **50:**1219-1229.

Müller D, Schopp P, Melchinger AE: **Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection.** *G3: Genes|Genomes|Genetics* 2018, **8:**1173.

Nakaya A, Isobe SN: **Will genomic selection be a practical method for plant breeding?** *Annals of Botany* 2012, **110:**1303-1316.

Nelson RL: **Managing self-pollinated germplasm collections to maximize utilization.** *Plant genetic resources: characterization and utilization* 2011, **9:**123-133.

Obolski U, Lewin-Epstein O, Even-Tov E, Ram Y, Hadany L: **With a little help from my friends: cooperation can accelerate the rate of adaptive valley crossing.** *BMC Evolutionary Biology* 2017, **17:**143.

Oddi FJ, Miguez FE, Ghermandi L, Bianchi LO, Garibaldi LA: **A nonlinear mixed-effects modeling approach for ecological data: Using temporal dynamics of vegetation moisture as an example.** *Ecology and evolution* 2019, **9:**10225-10240.

Pinheiro JC: *Mixed-effects models in S and S-PLUS / José C. Pinheiro, Douglas M. Bates.* New York: New York: Springer; 2000.

Pinheiro JC, Bates DJ, D. DebRoy S, Sakar D: *The Nlme Package: Linear and Nonlinear Mixed Effects Models, R Version 3.* 2012.

Podlich DW, Cooper M: **Modelling Plant Breeding Programs as Search Strategies on a Complex Response Surface.** In *Simulated Evolution and Learning: Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL'98 Canberra, Australia, November 24–27, 1998 Selected Papers.* Edited by McKay B, Yao X, Newton CS, Kim J-H, Furuhashi T. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999: 171-178

Pryce JE, Hayes BJ, Goddard ME: **Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information.** *Journal of Dairy Science* 2012, **95:**377-388.

Ramasubramanian V and Beavis WD. "Factors Affecting Response to Recurrent Genomic Selection in Soybeans." *bioRxiv* (2020): 2020.02.14.949008.

Rardin, Ronald L. *Optimization in Operations Research / Ronald L. Rardin, Purdue University*. Second edition.. ed.: Boston : Pearson, 2017.

Ryman N, Leimar O: **GST is still a useful measure of genetic differentiation — a comment on Jost's D.** *Molecular Ecology* 2009, **18:**2084-2087.

Saeki Y, Tudari, Crowley PH: **Allocation tradeoffs and life histories: a conceptual and graphical framework. (Report).** 2014, **123:**786.

Schierenbeck S, Pimentel ECG, Tietze M, Körte J, Reents R, Reinhardt F, Simianer H, König S: **Controlling inbreeding and maximizing genetic gain using semi-definite programming with pedigree-based and genomic relationships.** *J Dairy Sci* 2011, **94:**6143-6152.

Seada H, Deb K: **Non-dominated sorting based multi/many-objective optimization: Two decades of research and application.** In *Multi-Objective Optimization.* Springer; 2018: 1-24

Sheftel H, Shoval O, Mayo A, Alon U: **The geometry of the Pareto front in biological phenotype space.** *Ecology and Evolution* 2013, **3:**1471-1483.

Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U: **Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space.** *Science (New York, NY)* 2012, **336:**1157.

Skolicki Z, Jong KD: **The importance of a two-level perspective for island model design.** In *2007 IEEE Congress on Evolutionary Computation*; *25-28 Sept. 2007*. 2007: 4623-4630.

Skolicki ZM: **An analysis of island models in evolutionary computation.** George Mason University, 2007.

Sonesson A, Woolliams J, Meuwissen T: **Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection.** In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, 1*. 2010

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB: **Fingerprinting Soybean Germplasm and Its Utility in Genomic Research.** *G3: Genes|Genomes|Genetics* 2015, **5:**1999.

Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, Song B-H, Charles An Y-Q, Hyten D, Nelson R, et al: **Genetic Characterization of the Soybean Nested Association Mapping Population.** *The Plant Genome* 2017, **10**.

Specht JE, Diers BW, Nelson RL, de Toledo JFF, Torrion JA, Grassini P: **Soybean.** *Yield gains in major US field crops* 2014, **33:**311-355.

USDA-ERS, 2020 https://www.ers.usda.gov/data-products/commodity-costs-and-returns/commodity-costs-and-returns/#Recent%20Cost%20and%20Returns)

Whitley D, Rana S, Heckendorn RB: **The island model genetic algorithm: On separability, population size and convergence.** *CIT Journal of computing and information technology* 1999, **7:**33-47.

Woolliams JA, Berg P, Dagnachew BS, Meuwissen TH: **Genetic contributions and their optimization.** *J Anim Breed Genet* 2015, **132:**89-99.

Wray, N. R., and M. E. Goddard. "Increasing Long-Term Response to Selection." *Genetics Selection Evolution* 26, no. 5 (1994/10/15 1994): 431.

Xavier A, Muir WM, Rainey KM: **Assessing Predictive Properties of Genome-Wide Selection in Soybeans.** *G3: Genes|Genomes|Genetics* 2016, **6:**2611.

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan P, et al: **Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population.** *G3: Genes|Genomes|Genetics* 2017.

Xavier A: **Efficient Estimation of Marker Effects in Plant Breeding.** *G3: Genes|Genomes|Genetics* 2019, **9:**3855.

Yabe, S., M. Yamasaki, K. Ebana, T. Hayashi, and H. Iwata. "Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops." *PLoS One* 11, no. 4 (2016): e0153945.

Yv, Yv Haimes, L. Lasdon, and Da Wismer Da. "On a Bicriterion Formation of the Problems of Integrated System Identification and System Optimization." *IEEE Transactions on Systems, Man, and Cybernetics* (1971): 296-97.

Zadeh L: **Optimality and non-scalar-valued performance criteria.** *TAC* 1963, **8:**59-60.

Zuur A: ***Mixed Effects Models and Extensions in Ecology with R*** *by Alain Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith.* 1st ed. 2009.. edn: New York, NY : Springer New York : Imprint: Springer; 2009.

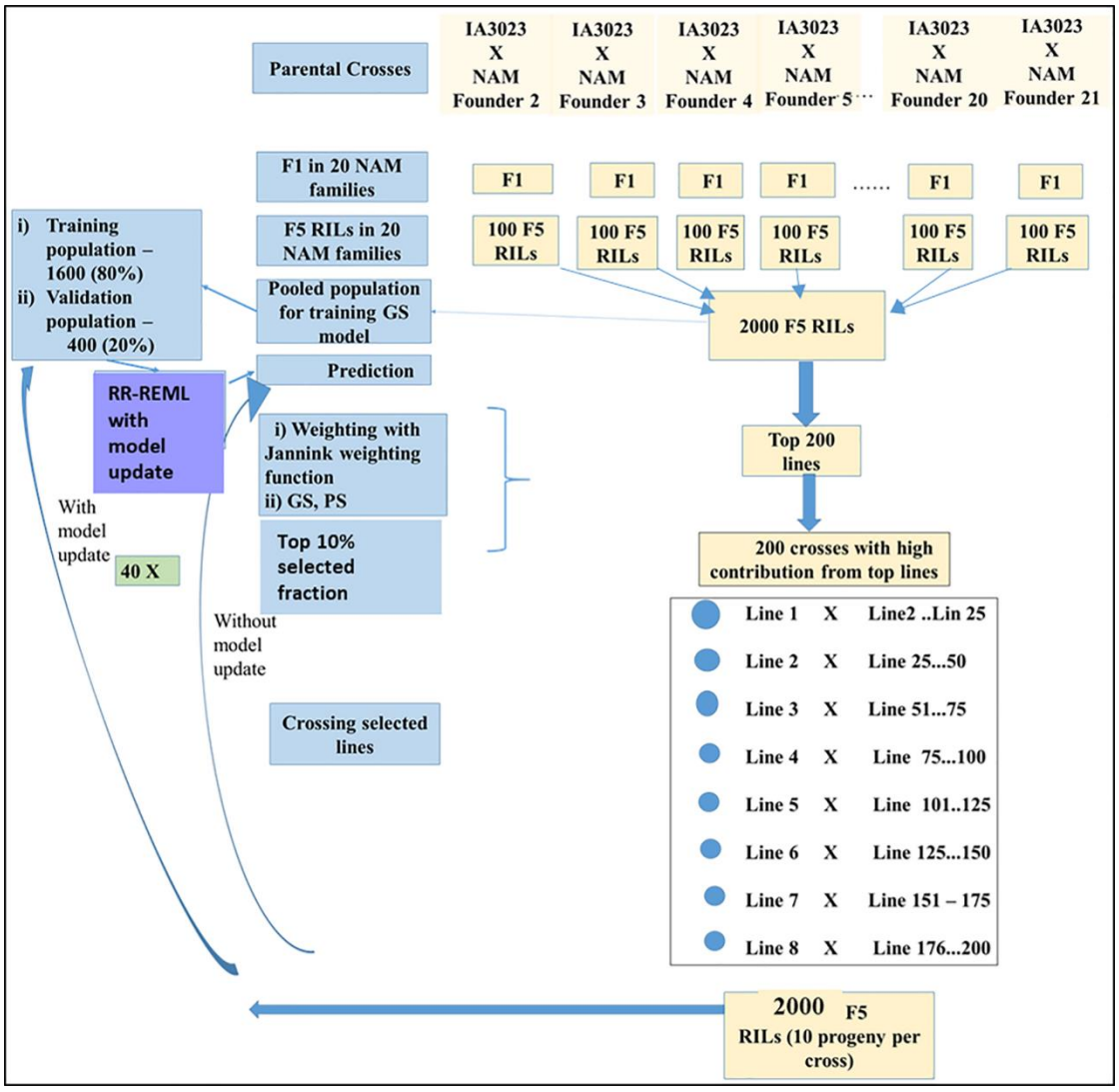

Figure 1 Schematic representing simulated recurrent selection in Non-Isolated (NI) populations comprised of twenty families from Soybean NAM founders. The schematic depicts the *in silico* steps used to generate the base population of 2000 F_5 RILs derived from 20 NAM founder lines crossed to IA3023. The depiction includes the model training step and the recurrent steps of prediction, sorting, truncation selection, crossing, and generation of 2000 F_5 RILs for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process should be continued for another cycle.

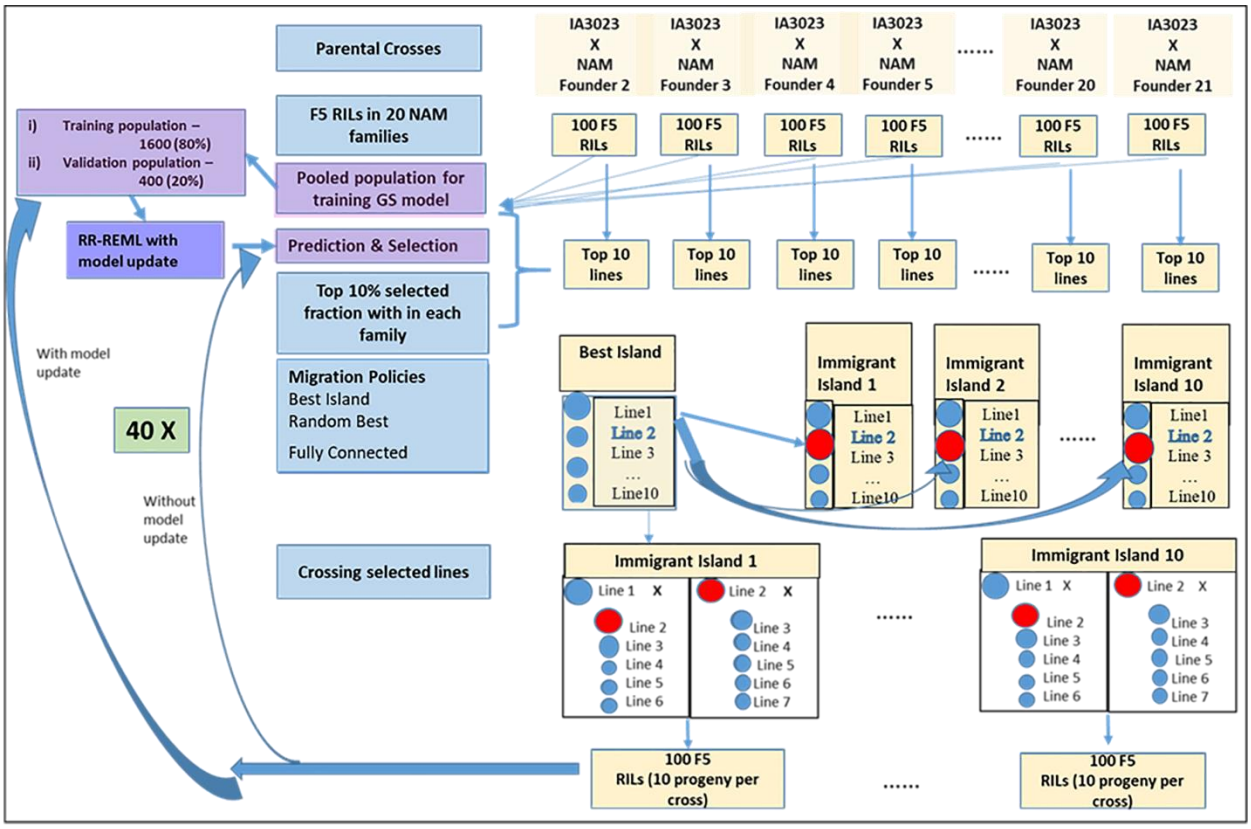

Figure 2 Schematic representing simulated recurrent selection of Family Island (FI) populations where each of the twenty families from the Soybean NAM founders is considered an island population. The schematic depicts the *in silico* steps used to generate the base population of 2000 F_5 RILs derived from 20 NAM founder lines crossed to IA3023. 100 F5 RILs generated from each of the crosses form a distinct island. The depiction includes the model training step and the recurrent steps of prediction, sorting, truncation selection within islands, migration, crossing, and generation of 200 F_5 RILs per island for each cycle as well as the decision steps to check if the training set should be updated and if the recurrent process should be continued for another cycle. The blue shaded circles represent lines that are descendants of the founder populations in the islands and red shaded circles represent lines that are replaced by immigrants from the island with the largest genotypic value for the 'Best Island' policy.

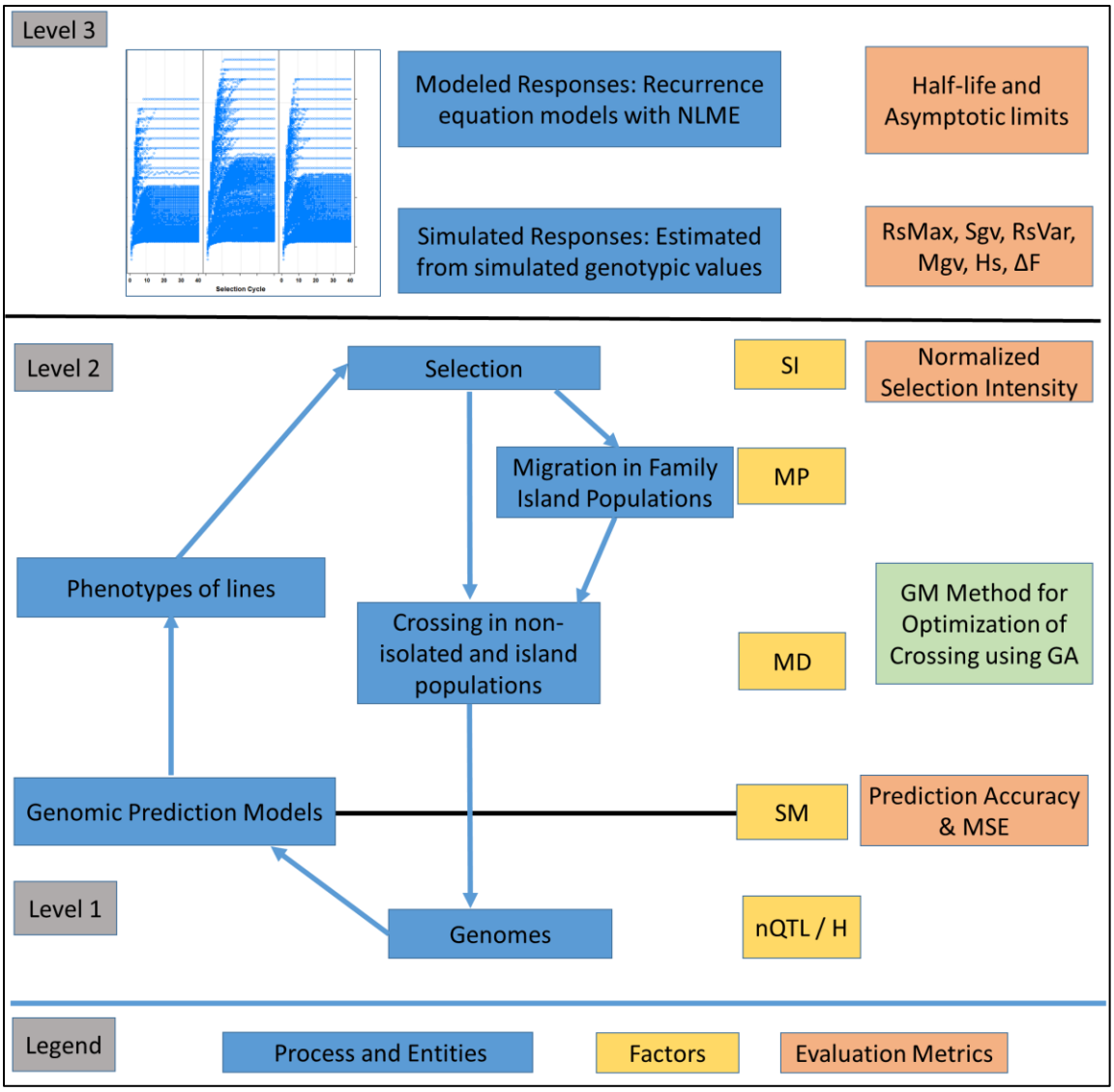

Figure 3 Overview of the Recurrent Selection Process: Representation of entities such as genomes, associated RILs and processes such as the estimation of marker effects, selection, migration and crossing. Levels correspond to layers of information with level 1 comprised of genomic information, level 2 comprised of phenotypes of lines within and across families and level 3 comprised of higher level information including responses across cycles of selection. The factors include nQTL and H at the genome level, SM (selection method) including PS, GS and WGS. The factors at level 2 includes SI (top 10% selected fraction), MD (Mating design, which includes Hub Network, Chain Rule, Random Mating, and Genomic mating levels) and MP (Migration Policy, which includes "Best Island", "Random Best" and "Fully Connected" policies). Among the BD levels, GM method involves application of

evolutionary multi-objective optimization to minimize inbreeding and maximizing gain. Level 3 is characterized using evaluation metrics such as half-life and asymptotic limits derived from recurrence equation models and metrics such as standardized responses (Rs), Standardized genetic variance (Sgv), Maximal genotypic values (Mgv) and efficiency of converting loss in genetic variance into gain (RsVar) derived from simulated outcomes. Other metrics include prediction accuracy and MSE for GP models (RR-REML) and expected heterozygosity (Hs).

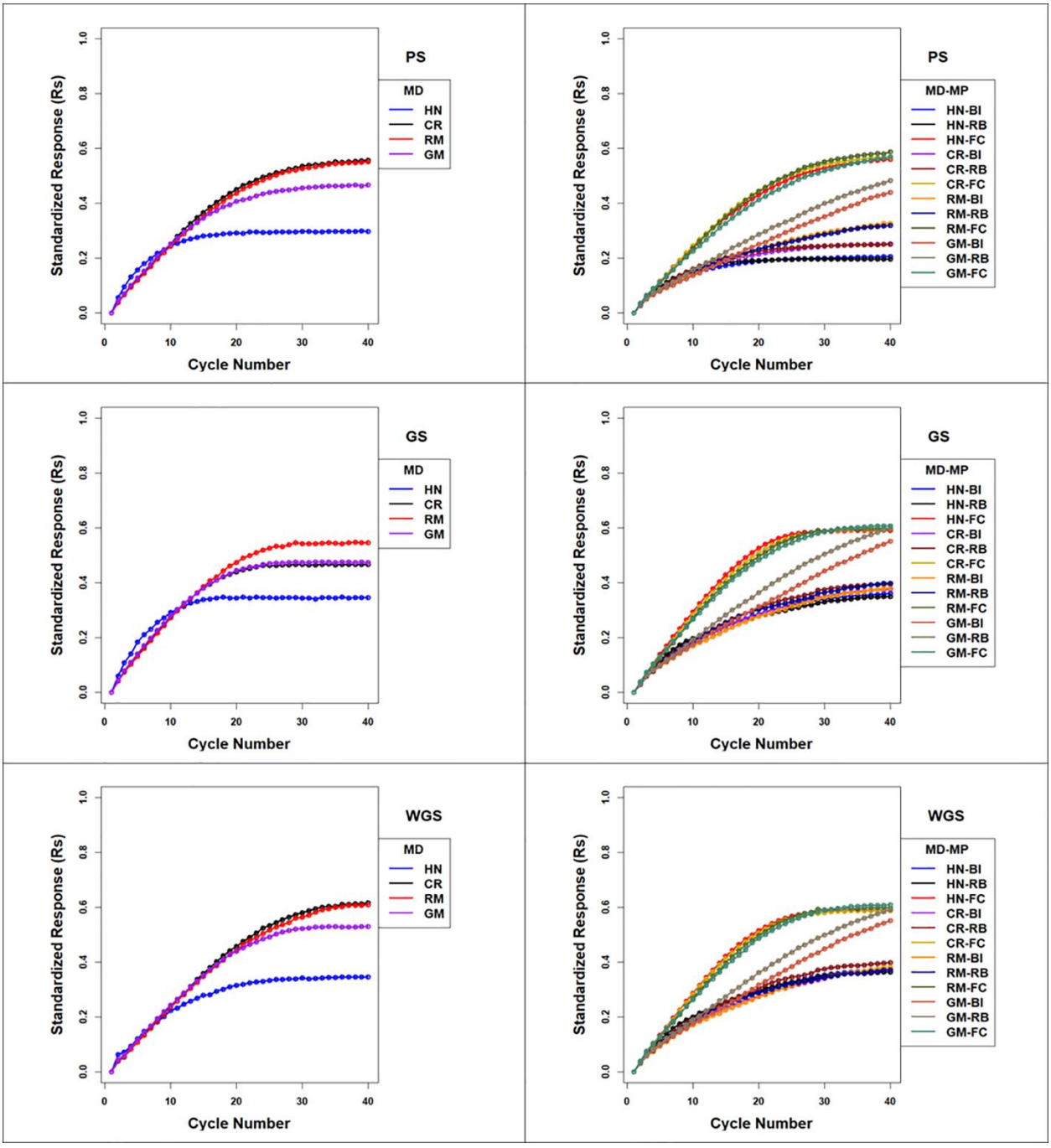

Figure 4. Standardized Genotypic Responses (Rs) across 40 cycles of recurrent selection on non-isolated (left panels) and family island (right panels) populations, using PS (top panels), GS (middle panels) and WGS (bottom panels) for the four mating designs: Hub-Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM). Standardized genotypic responses are represented from a simulated genetic architecture consisting of 400 additive QTL uniformly distributed throughout the genome and responsible for 70% of phenotypic variability. Ten percent of lines are selected to be used in crosses in

HN, CR, RM and GM designs. Migration policies include bi-directional migrations of two migrants every other cycle involving the Best Island (BI), Random Best (RB), and Fully Connected (FC) migration policies.

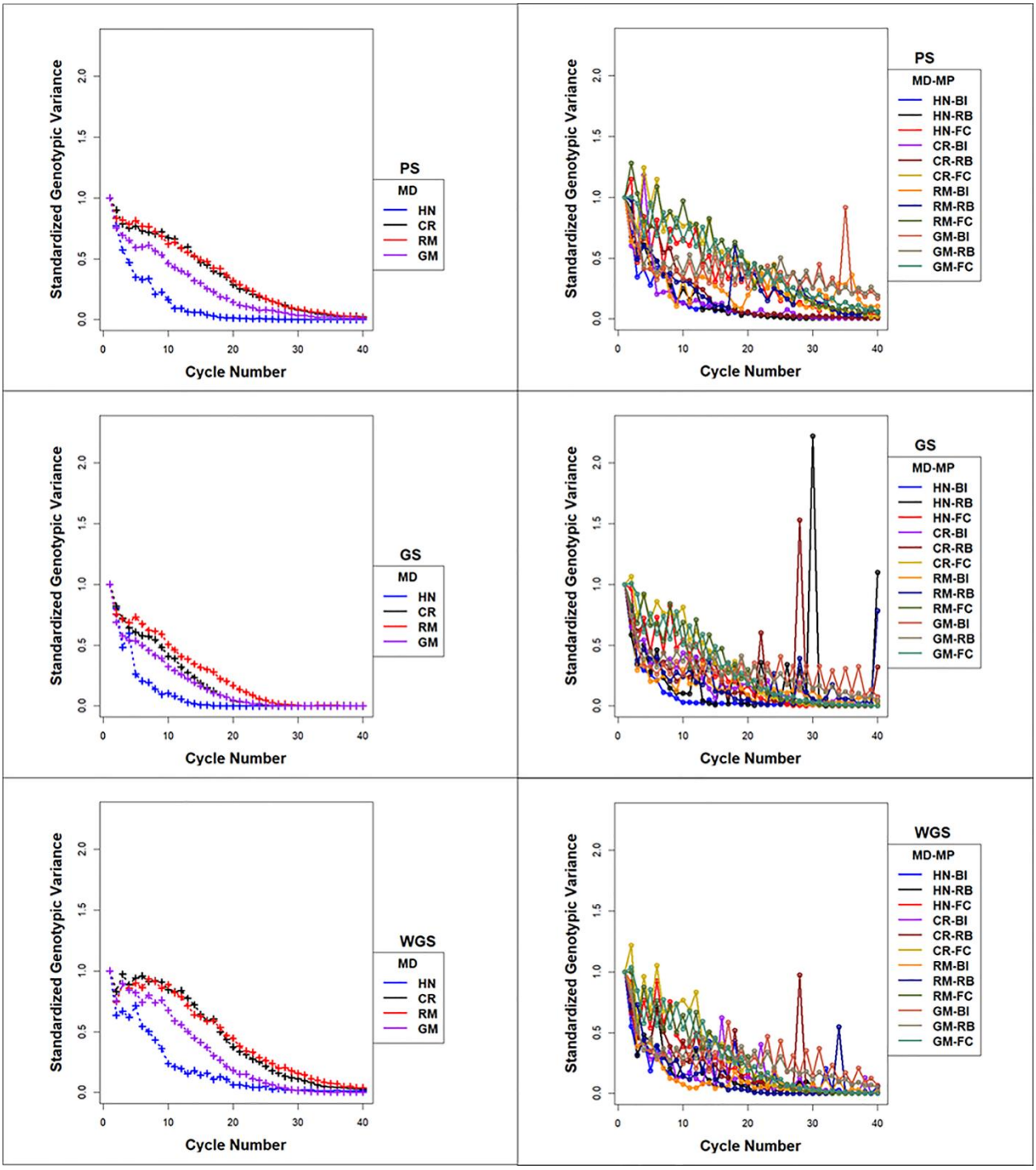

Figure 5 Standardized Genotypic Variance across 40 cycles of recurrent selection on non-isolated (left panels) and family island (right panels) populations, using PS (top panels), GS (middle panels) and WGS (bottom panels) for the four mating designs: Hub-Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM). Ten percent of lines are selected for crossing. The genetic architecture in the initial simulated founder lines

consisted of 400 additive QTL uniformly distributed throughout the genome and expressed broad sense heritability on an entry mean basis of 0.7. Genetic variance is standardized to the average genotypic variance in founder populations in cycle ‘0’. Average island genetic variance refers to genetic variance within families averaged across 20 families. Migration policies in the island models consisted of bidirectional exchange of two immigrants and emigrants every other cycle of selection. Migration policies include BI- “Best Island”, RB- “Random Best”, and FC- “Fully Connected”

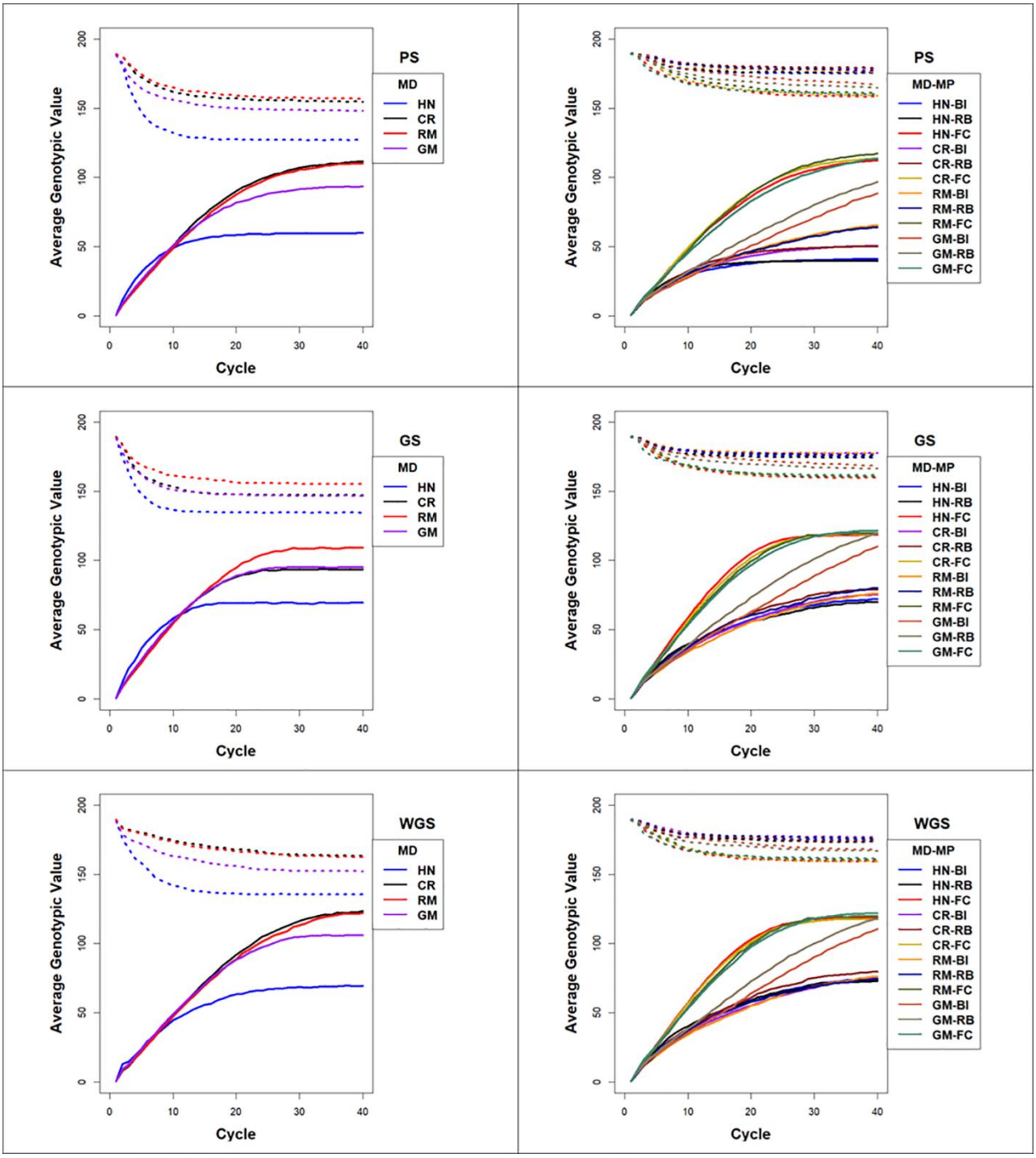

Figure 6 Lost Genotypic Potential and Average Genotypic Values across 40 cycles of recurrent selection on non-isolated (left panels) and island (right panels) populations, using Phenotypic Selection (PS-top row of panels), Genomic Selection (GS-middle row of panels) and Weighted Genomic Selection (WGS-bottom row of panels) and four mating designs: Hub-Network (HN), Chain Rule (CR), Random Mating (RM), and Genomic Mating (GM).

Ten percent of lines are selected for crosses using HN, CR, RM and GM mating designs. The genetic architecture in the initial simulated founder lines consisted of 400 additive QTL uniformly distributed throughout the genome and expressed broad sense heritability on an entry mean basis of 0.7. The dotted lines represent maximum genetic potential estimated from favorable alleles that are lost from the population and solid lines represent increase in average genotypic value of populations due to recurrent selection. Migration policies in the island models consisted of bidirectional exchange of two immigrants and emigrants every other cycle of selection. Migration policies include BI- “Best Island”, RB- “Random Best”, and FC- “Fully Connected”

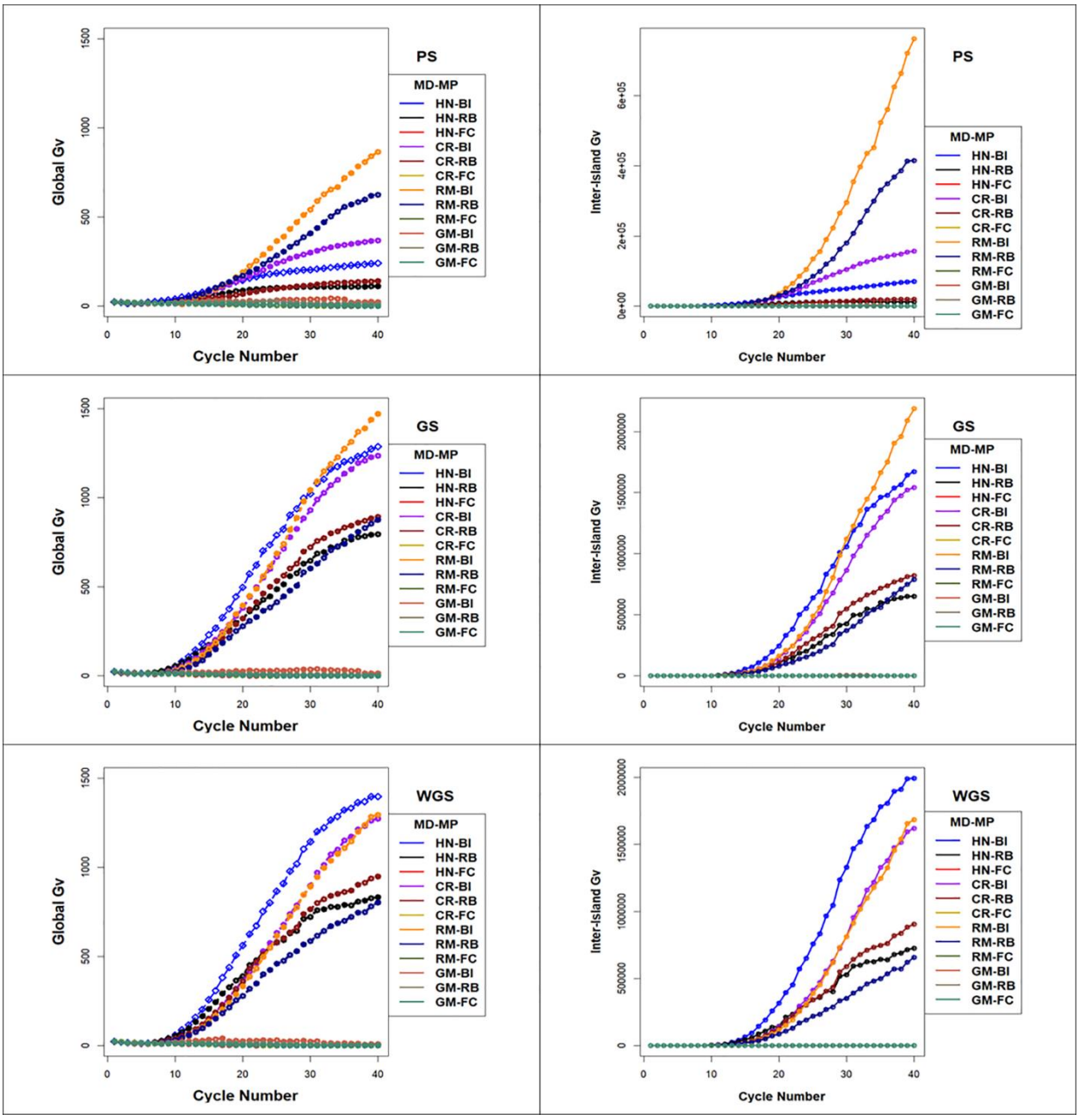

Figure 7. Global and Inter-Island Genotypic Variance in Island Selection i) **G**lobal genotypic variance (Left Panel), and ii) Inter-island genetic variance (Right Panel) for PS (top), GS (middle) and WGS (bottom) for the four mating designs including HN, CR, RM and GM methods and three migration policies including BI, RB and FC for 400 simulated QTL and 0.7 H. Genotypic variance is standardized to the average genotypic variance in founder

populations in cycle ‘0’. GP models are updated every cycle in GS and WGS using training sets with data from all prior cycles of selection. PS- Phenotypic Selection, GS- Genomic Selection, WGS -Weighted Genomic Selection. Mating Design: HN (Hub Network), CR (Chain rule), and RM (Random Mating), GM (Genomic Mating) method. Migration policies in the island models consisted of bidirectional exchange of two immigrants and emigrants every other cycle of selection. Migration policies include BI- “Best Island”, RB- “Random Best”, and FC- “Fully Connected”

Table 1 Treatment Design representing the factors that impact responses and limits of responses. The parameter values for levels of island selection specific factors were selected based on limits of responses from a larger set of simulations (2664 combinations of factors with 10 replicates per combination of factor) with migration frequency (1, 2, 3), migration size (1, 2) and migration direction (1, 2) and mating designs (HN, CR and RM) for 40, 400 and 4289QTL with 0.7 and 0.3 H.

Factors	Levels	Values for Levels
Population Type	2	Non-isolated and Island populations
Island Model Selection		
Migration Frequency	1	Migration frequency of 2 corresponds to migration of lines every other cycle of selection
Migration Size	1	Migration of 2 lines per migration event (20%).
Migration Policy	4	i) Discrete Selection (For DS, migration frequency, size and direction are set to '0') ii) Best Island iii) Random Best iv) Fully connected
Migration Direction	1	i) Bi-directional
Factors Common to Non-isolated and Island Populations		
Selection Method	3	i) PS ,ii) GS, iii) WGS
Mating Design	4	i) Hub Network Design ii) Chain Rule iii) Random mating iv) Genomic mating
Genetic Model Parameters	1	400 QTL and 0.7 H
Total Number of Combinations of Treatment Factors	60	
Total Number of Simulations	**5 (replicates/combination of factors)**	**300**

Table 2 Trade-off Table for Strategies Tradeoff table to support decision for selecting the best strategy to achieve possible objectives including maximum gain in 5, 10, 20, 30 and 40 cycles of recurrent selection. The methods are ranked for each of the objectives based on the absolute genotypic response values for each of the methods are provided along with the ranking of the method for the specific objective within parenthesis. Three objective weights are provided to define the relative importance of the objectives: i) the weighted rank of methods are estimated with more emphasis on the first 20 cycles (top), ii) the weighted rank of methods are estimated with equal emphasis on the first and last 20 cycles (bottom). The best five methods among the 60 methods for each of the weighted objectives are presented. The simulations are provided for 400 simulated QTL responsible for 70% of phenotypic variability. Migration policies include "Discrete Selection", "Best Island", "Random Best", and "Fully Connected". Other migration factors correspond to migration frequency -2, migration direction -2 (bi-directional), and migration size - 2. Selection methods include PS-Phenotypic Selection, GS- Genomic Selection, and WGS-Weighted Genomic Selection. Mating Design includes HN (Hub Network), CR (Chain rule), RM- Random Mating, GM- Genomic Mating method.

Objectives	**Objective Weights for gain in the first 20 cycles**	**Method**				
		IM-FC-GS-HN	**IM-FC-WGS-HN**	**IM-FC-GS-CR**	**NI-GS-CR**	**NI-GS-GM**
Rs in 5 cycles (Rank)	**0.5**	0.14 (5)	0.13(13)	0.13 (13)	0.14 (5)	-0.14 (5)
Rs in 10 cycles (Rank)	**0.2**	0.29 (3)	0.29 (3)	0.28 (7)	0.28 (7)	0.28 (7)
Rs in 20 cycles (Rank)	**0.1**	0.53 (1)	0.51 (4)	0.51 (4)	0.44 (18)	0.44 (18)
Rs in 30 cycles (Rank)	**0.1**	0.59 (7)	0.59 (7)	0.59 (7)	0.47 (22)	0.47 (22)
Rs in 40 cycles (Rank)	**0.1**	0.59 (13)	0.59 (13)	0.60 (8)	0.47 (26)	0.47 (26)
Weighted Rank		**1**	**2**	**3**	**4**	**4**
Objectives	**Objective weighted equally for gain across the 40 cycles**	**Method**				
		IM-FC-GS-GM	**IM-FC-WGS-GM**	**IM-FC-GS-CR**	**IM-FC-WGS-RM**	**NI-WGS-CR**

Table 2. Continued

		IM-FC-GS-GM	IM-FC-WGS-GM	IM-FC-GS-CR	IM-FC-WGS-RM	NI-WGS-CR
Rs in 5 cycles / Rs	**0.2**	0.13(13)	0.13(13)	0.13 (13)	0.13 (13)	-0.11 (29)
Rs in 10 cycles / Rs	**0.2**	0.27 (11)	0.26 (12)	0.28 (7)	0.27 (11)	0.24 (22)
Rs in 20 cycles / Rs	**0.2**	0.48 (8)	0.49 (7)	0.51 (4)	0.5 (6)	0.46 (10)
Rs in 30 cycles / Rs	**0.2**	0.59 (7)	0.59 (7)	0.59 (7)	0.59 (7)	0.58 (9)
Rs in 40 cycles / Rs	**0.2**	0.61 (4)	0.61 (4)	0.6 (8)	0.6 (8)	0.62 (1)
Weighted Rank		**1**	**1**	**3**	**4**	**4**

CHAPTER 4. GENERAL CONCLUSION

4.1 Summary

In the past two decades the use of whole genome information prompted the development of analytic tools for genetic improvement (Bernardo and Yu 2007; Bernardo 2008; Bevan et al. 2017; Hickey et al. 2017; Andorf et al. 2019; Mascher et al. 2019). These new tools led to a large number of studies to examine the potential of methods such as genomic selection in crop improvement projects (Andorf et al. 2019). Novel approaches, however need to be supported by studies on the impact of genomic selection methods on rates and limits of responses to selection as well as understanding tradeoffs between short term and long term breeding objectives.

In order to optimize tradeoffs between rates of genetic gains and retention of useful genetic variance in soybean, we evaluated the impact of prediction models and selection strategies on response to selection across 40 cycles of recurrent selection. While soybean breeders currently take four to five years to complete a cycle of recurrent selection, emerging technologies will enable soybean breeders to complete a cycle in a year. Thus, our work will serve as a comparator for future investigations on the impact of rapid cycling in soybean. By decomposing breeding strategies into combinations of population structure, selection methods, mating designs and migration policies we demonstrated potential responses to recurrent selection for a small soybean genetic improvement project using contemporary soybean germplasm adapted to MZ's II and III. In chapter 2, we introduced the use of first order recurrence equation as a tool for analyses of recurrent selection. The tool enabled us to report the impact of training genomic prediction models and training sets on short-term response rates, populations' half-lives and long-term limits to selection. We also reported the impact of selection intensity, number of QTL and heritability on genotypic responses. In chapter 3, we evaluated the rates and limits of

response to selection with several breeding strategies and examined the application of algorithms developed in the field of evolutionary computing and genetic algorithms to maximize retention of useful genetic variability with minimal loss to rates of short-term genetic gains.

We adopted the perspective that genetic improvement projects can be regarded as either single or multiple connected search strategies in genotypic space (Podlich and Cooper 1999; Cooper et al. 2002; Cooper et al. 2014). The single search strategy, a.k.a. global search strategy, corresponds to selection of lines among non-isolated families. The multiple connected search strategy, a.k.a. local search strategy, corresponds to selection of lines in multiple domains with exchange of lines among domains. In addition to the perspective of global or local search strategies, selection can be viewed as cooperation vs. competition and exploitation vs. exploration. By characterizing selection methods, mating designs and migration policies for their levels of cooperation or exploration, we demonstrated that it is possible to tune parameters that control relative levels of cooperation or exploration in global or local search strategies and as a consequence to find many options that can be used to meet the breeding objectives of any specific genetic improvement program. From the perspective of tradeoffs between exploration and exploitation, local selection among and within islands enables exploration of diverse domains resulting in greater probabilities of finding solutions with greater genotypic values (higher peaks or limits at convergence).

By treating families as islands within a project, we opened the opportunity to consider soybean genetic improvement projects across MZ's II and III as islands where genotypes can be exchanged among project islands based on annual evaluations in uniform regional trials and according to legal licensing rules. Currently soybean breeding projects exchange only the best performing lines adapted to the same MZ and similar climatic conditions. Our results suggest

that this *ad hoc* immigration policy is not optimal, none-the-less it enables soybean breeders to slow erosion of useful genetic variability by exchanging contemporary germplasm among projects. An advantage of island selection is that diversity among islands increased across 40 cycles of recurrent selection, even as diversity within islands decreased. Beyond 40 cycles of recurrent selection, we expect genetic variability among islands will decrease.

We demonstrated that simulations can be useful to bridge the gap between theoretical studies of evolutionary computing and genetic improvement in soybean breeding programs. Of course, not all of the factors easily examined in theoretical models can be implemented in real PB programs. For example, while it is relatively easy to simulate factors that account for how frequently and how many lines to exchange among islands, it is difficult to include these factors in the decision making process. While it is easy to learn about the general characteristics of strategies from simulations, it is difficult to recommend any strategies for any specific breeding program without clear articulation of the breeding objectives in terms of measurable metrics. It is clear that most commercial soybean breeding programs will hesitate to compromise immediate genetic gains for retention of useful genetic diversity. Should retention of useful genetic variability be of greater emphasis in public soybean breeding programs? How much greater emphasis? Even clearly articulated measurable objectives will require investments to build better models that take into account complexities of multiple operational decisions associated with annual activities and subsequently develop search algorithms that can find optimal solutions for the competing objectives in dynamic environmental and economic landscapes (Byrum et al, 2017; Andorf et al. 2019; Li et al. 2021).

As a first step toward a more comprehensive decision support system, it will be useful to reframe analyses of genetic potential, as well as rates and limits of genetic gains, in a Cost, Time

and Probability of Success (CTP) framework. Several researchers in our group have framed the optimization of trait introgression projects in a CTP framework (Cameron et al. 2017; Han et al. 2017). Applying GS for replenishing genetic potential lost due to selection and integration with methods such as genomic mating and OCS (Akdemir and Sanchez 2016; Gorjanc et al. 2018) is a promising research direction. By framing success as restoring a defined proportion of genetic potential from other active breeding islands or archived populations, it should be possible to to define combinations of factors that will assure selection of lines for crossing and migration that will be optimal with respect to any specific project objectives.

4.2 Archived Islands of Genetic Diversity.

Germplasm collections for many agricultural crops include large numbers of genotypes from geographical regions around the globe. Soybean breeders and geneticists agree that the germplasm collection holds useful genetic variability that could contribute to genetic gains for many agronomic traits (Yu X et al. 2016; Xavier et al, 2018). However, only small numbers of the collections have been utilized in applied plant breeding programs. For example, only 1000 out of the 45000 unique accessions of Soybean available in the USDA soybean germplasm collection are used in applied plant breeding (Jarquin et al, 2016; Xavier et al, 2018). The USDA germplasm collection can be clustered into subpopulations based on their distinct collection sites and morphological characteristics. The lines within a cluster often share similar phenotypic characteristics for important agronomic traits and are often adapted to specific geographical regions (Xavier et al, 2018). Song et al (2015) genotyped 19,648 accessions of the USDA soybean germplasm collection using the Infinium II Bead chip consisting of 52,041 SNPs .The assayed accessions include cultivated varieties, land races and wild type soybean from 84

countries. (Song et al. 2015). The majority of these accessions form distinct clusters based on their country of origin (Song et al. 2015; Bandillo et al. 2015; Jarquin D et al. 2016; Xavier et al. 2018).

Analysis of SoySNP50K data for 14,430 unique accessions (after removal of duplicate and isogenic lines) from the collection with ADMIXTURE (Alexander et al. 2009) for k=5 resulted in population clusters that were highly correlated with geographical origin. Whereas applying agglomerative clustering with Ward's agglomerative method and Nei's genetic distance resulted in 8 clusters, each consisting of accessions from the same geographical region cluster-specific phenotypic characteristics (Xavier et al. 2018). A GWAS study on members representing each cluster revealed some marker trait associations (MTAs) in genomic regions that are common to all subpopulations and in other genomic regions that were unique to some subpopulations (Bandillo et al. 2015; Xavier et al. 2018). This shows that favorable alleles are distributed unequally among sub-populations and also implies that these alleles are distributed unequally among and within linkage groups. In our simulations, we assumed equal contributions of positive alleles from all founders with the QTL distributed uniformly among and within linkage groups. Future simulations will needs to investigate more realistic scenarios with unequal distributions. Identification and migration of lines that carry favorable haplotype blocks instead of favorable alleles might turn out to be a better strategy.

Selecting genotypes for crossing from a large collection assumes that evaluation of accessions *per se* will be directly useful in existing breeding populations. Such selection does not consider how the new sources of favorable alleles will alter the existing genetic landscape composed of varieties created during the last 80 years (~ eight cycles) of recurrent selection. An example of how germplasm can be evaluated in the context of existing genetic landscape is the

NCSRP supported SoyNAM project. The purpose of SoyNAM was to investigate potential for novel useful genetic contributions to agronomic traits from samples of germplasm adapted to Maturity Zones II to IV (Song et al. 2017). The sampled germplasm included contemporary public varieties, plant introduction accessions and lines from the USDA soybean germplasm conversion program. The SoyNAM panel was generated from crossing 40 founder lines with a common parental inbred line, IA3023. F1 individuals from each of the 40 crosses were selfed for five generations to generate 140 Recombinant Inbred Lines (RILs). The 40 founder lines consisted of 17 high-yielding elite cultivars (EL) developed in the mid-western US, 15 adapted breeding lines (BX) of diverse ancestry developed at USDA-ARS, and 8 plant introductions (PI) with exotic ancestry including lines developed in South Korea, Russia and China (Song et al. 2017; Diers et al. 2018). The parental lines and the 5600 RILs were grown in 22 environments across Mid-Western US from 2011 to 2013 and phenotyped for several important agronomic traits (Diers et al. 2018). Genotypic data from SoyNAM 6K bead chip and data for the founder lines from the SoyNAM 50K bead chip are available (Song et al. 2015; Song et al. 2017; Diers et al. 2018). Results indicate that most identified favorable alleles for agronomically important traits exist in the contemporary varieties, although there are some notable exceptions from the BX genotypes. Not only are favorable alleles unequally distributed among the germplasm resources, in the context of the existing selected breeding populations, positive contributions will require some sort of breeding bridge (sometimes referred to as pre-breeding) such as identified in the lines created by the BX project.

4.3 Future Research

In breeding populations consisting of genetic diversity similar to the SoyNAM panel, we found that application of Island Models (IM) with Genomic Mating (GM) provided a range of trade-offs between the competing goals of retaining useful genetic variance while assuring competitive genetic gains. In some cases rates of genetic gain in early cycles were almost as great as the best rates and the greatest proportion of genetic potential among the founders was realized in later cycles. We are not aware of research to identify the best combinations of selection and crossing methods for retrieving lost favorable alleles from recently archived populations. Analysis of exPVP genotypic data in soybean and corn reveal distinct clusters of lines that are associated with year of release and location of development (Beckett et al. 2017). Such clusters in germplasm resources have unequal distribution of favorable alleles, many of which are likely in linkage disequilibrium with unfavorable alleles. The clusters also have different numbers of members, with the elite PVP's often grouped in small clusters, whereas the clusters comprised of lines in broader germplasm resources often have large membership. Such unequal distribution of membership per cluster further complicates implementation of island model based optimization. It may be that maintaining different objectives for different types of islands will improves global diversity. Elite varieties have relatively shorter life spans and are often archived after 5-10 years. Thus, some archived lines are still protected by PVP certificates. Other lines that have expired PVP certificates after ~20 years of protection and founder lines that were developed 20-40 years ago form archived islands that are not as far removed from active breeding populations as germplasm collections. These lines could become part of germplasm conversion islands with objectives to rapidly break unfavorable linkage blocks for better access to favorable alleles.

While there have been many proposals to apply genome editing tools for sustainable crop improvement, the use of genome editing is still in its infancy (Bevan et al. 2017; Hickey et al. 2017; Andorf et al. 2019; Mascher et al. 2019). Given that the predominant model for quantitative traits is the infinitesimal model consisting of many loci with small additive effects contributing to the trait value, such gene editing approaches will be ideally combined with approaches such as genomic selection; for example, see Promotion of Alleles by Gene Editing (PAGE) (Janez et al. 2015; Hickey et al. 2016, 2017). We speculate that there may be a need for islands of breeding, where genome edited varieties that preserve crop diversity are maintained and developed. Costs for maintaining such islands need to be included when investigating their potential. But, just as costs need to be included, so should benefits because the benefits could offset losses by serving as a resource pool for replenishing lost useful variance in breeding populations. Island models could be applied in such contexts for optimization of processes to achieve objectives set by the program.

Other approaches with potential for step changes in crop improvement pipeline include high-throughput phenotyping using high-resolution digital technologies, precision envirotyping, and integration with crop models and selection on network of loci or genes that function in gene regulatory and metabolic networks (Heslot et al. 2014; Andorf et al. 2019; Diego et al. 2021; Li et al. 2021). As these innovative approaches for crop improvement are proposed, we hope that they will be evaluated in the context of optimizing breeding objectives so that the ideal responses to selection can be approached.

References

Akdemir D, Beavis W, Fritsche-Neto R, Singh AK, Isidro-Sánchez J: **Multi-objective optimized genomic breeding strategies for sustainable food improvement.** *Heredity* 2019, **122:**672-683.

Akdemir D, Sánchez JI: **Efficient Breeding by Genomic Mating.** *Frontiers in Genetics* 2016, **7**.

Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome research* 2009, **19:**1655-1664.

Allier A, Lehermeier C, Charcosset A, Moreau L, Teyssèdre S: **Improving Short- and Long-Term Genetic Gain by Accounting for Within-Family Variance in Optimal Cross-Selection.** *Frontiers in Genetics* 2019, **10**.

Allier A, Moreau L, Charcosset A, Teyssèdre S, Lehermeier C: **Usefulness Criterion and Post-selection Parental Contributions in Multi-parental Crosses: Application to Polygenic Trait Introgression.** *G3: Genes|Genomes|Genetics* 2019, **9:**1469.

Andorf C, Beavis WD, Hufford M, Smith S, Suza WP, Wang K, Woodhouse M, Yu J, Lübberstedt T: **Technological advances in maize breeding: past, present and future.** *Theoretical and Applied Genetics* 2019, **132:**817-849.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A: *A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection.* 2015.

Beckett TJ, Morales AJ, Koehler KL, Rocheford TR: **Genetic relatedness of previously Plant-Variety-Protected commercial maize inbreds.** *PloS one* 2017, **12:**e0189277-e0189277.

Bernardo R: **Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years.** *Crop Science* 2008, **48:**1649.

Bernardo R, Yu J: **Prospects for Genomewide Selection for Quantitative Traits in Maize All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.** *Crop Science* 2007, **47:**1082-1090.

Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD: **Genomic innovation for crop improvement.** *Nature* 2017, **543:**346-354.

Byrum J, Beavis B, Davis C, Doonan G, Doubler T, Kaster V, Mowers R, Parry S: **Genetic Gain Performance Metric Accelerates Agricultural Productivity.** *Interfaces* 2017, **47:**442-453.

Cameron JN, Han Y, Wang L, Beavis WD: **Systematic design for trait introgression projects.** *Theoretical and Applied Genetics* 2017, **130:**1993-2004.

Carter T, Nelson R, Sneller C, Cui Z, Boerma H, Specht J: **Genetic diversity in soybean.** *Soybeans: Improvement, production, and uses* 2004, **3**.

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G: **Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction.** vol. 65. pp. 311-336. Melbourne; 2014:311-336.

Cooper M, Podlich D, Micallef K, Smith O, Jensen N, Chapman S, Kruger N: **Complexity, quantitative traits and plant breeding: a role for simulation modelling in the genetic improvement of crops.** *Quantitative genetics, genomics and plant breeding'(Ed MS Kang) pp* 2002**:**143-166.

Dempewolf H, Baute G, Anderson J, Kilian B, Smith C, Guarino L: **Past and Future Use of Wild Relatives in Crop Breeding.** *Crop Science* 2017, **57:**1070-1082.

Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, et al: **Genetic Architecture of Soybean Yield and Agronomic Traits.** *G3: Genes|Genomes|Genetics* 2018.

Gorjanc G, Gaynor RC, Hickey JM: **Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection.** *Theoretical and Applied Genetics* 2018, **131:**1953-1966.

Gorjanc G, Jenko J, Hearne SJ, Hickey JM: **Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations.(Report).** *BMC Genomics* 2016, **17**.

Han Y, Cameron JN, Wang L, Beavis WD: **The Predicted Cross Value for Genetic Introgression of Multiple Alleles.** *Genetics* 2017, **205:**1409.

Heslot N, Akdemir D, Sorrells ME, Jannink J-L: **Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions.** *Theoretical and Applied Genetics* 2014, **127:**463-480.

Hickey JM, Bruce C, Whitelaw A, Gorjanc G: **Promotion of alleles by genome editing in livestock breeding programmes.** *Journal of Animal Breeding & Genetics* 2016, **133:**83-84.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB: **Impacts of genetic bottlenecks on soybean genome diversity.** *Proc Natl Acad Sci U S A* 2006, **103:**16666-16671.

Jarquin D, de Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, et al: **Utility of Climatic Information via Combining Ability Models to Improve Genomic Prediction for Yield Within the Genomes to Fields Maize Project.** *Frontiers in Genetics* 2021, **11**.

Jarquin D, Specht J, Lorenz A: **Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions.** *G3: Genes|Genomes|Genetics* 2016, **6:**2329.

Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, Hickey JM: **Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs.** *Genetics Selection Evolution* 2015, **47:**55.

Langewisch T, Lenis J, Jiang G-L, Wang D, Pantalone V, Bilyeu K: **The development and use of a molecular model for soybean maturity groups.** *BMC Plant Biology* 2017, **17:**91.

Li X, Guo T, Wang J, Bekele W, Sukumaran S, Vanous AE, McNellie JP, Cortes LT, Lopes M, Lamkey KR, et al: **An Integrated Framework Reinstating the Environmental Dimension for GWAS and Genomic Selection in Crops.** *Molecular Plant* 2021.

Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N: **Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding.** *Nature Genetics* 2019, **51:**1076-1081.

Podlich DW, Cooper M: **Modelling Plant Breeding Programs as Search Strategies on a Complex Response Surface.** In *Simulated Evolution and Learning: Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL'98 Canberra, Australia, November 24–27, 1998 Selected Papers.* Edited by McKay B, Yao X, Newton CS, Kim J-H, Furuhashi T. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999: 171-178

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB: **Fingerprinting Soybean Germplasm and Its Utility in Genomic Research.** *G3: Genes|Genomes|Genetics* 2015, **5:**1999.

Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, Song B-H, Charles An Y-Q, Hyten D, Nelson R, et al: **Genetic Characterization of the Soybean Nested Association Mapping Population.** *The Plant Genome* 2017, **10**.

Xavier A, Thapa R, Muir WM, Rainey KM: **Population and quantitative genomic properties of the USDA soybean germplasm collection.** *Plant Genetic Resources* 2018**:**1-11.

Yabe S, Yamasaki M, Ebana K, Hayashi T, Iwata H: **Island-Model Genomic Selection for Long-Term Genetic Improvement of Autogamous Crops.** *PLoS One* 2016, **11:**e0153945.

Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE, Roozeboom KL, Wang D, Wang ML, Pederson GA, et al: **Genomic prediction contributing to a promising global strategy to turbocharge gene banks.** *Nature Plants* 2016, **2:**16150.